Educational Psychology Review (2023) 35:113
https://doi.org/10.1007/s10648-023-09823-4

META-ANALYSIS

Effects of Rubrics on Academic Performance, Self-Regulated Learning, and self-Efficacy: a Meta-analytic Review

Ernesto Panadero[1] · Anders Jonsson[2] · Leire Pinedo[3] · Belén Fernández-Castilla[4]

Accepted: 9 October 2023 / Published online: 7 December 2023

Abstract

Rubrics are widely used as instructional and learning instrument. Though they have been claimed to have positive effects on students' learning, these effects have not been meta-analyzed. Our aim was to synthesize the effects of rubrics on academic performance, self-regulated learning, and self-efficacy. The moderator effect of the following variables was also investigated: year of publication, gender, mean age, educational level, type of educational level (compulsory vs. higher education), number of sessions, number of assessment criteria, number of performance levels, use of self and peer assessment, research design, and empirical quality of the study. Standardized mean differences (for the three outcomes) and standardized mean changes (SMC; for academic performance) were calculated from the retrieved studies. After correcting for publication bias, a moderate and positive effect was found in favor of rubrics on academic performance ($g = 0.45$, $k = 21$, $m = 54$, 95% CI [0.312, 0.831]; SMC = 0.38, 95% CI [0.02, 0.75], $k = 12$, $m = 30$), whereas a small pooled effect was observed for self-regulated learning ($g = 0.23$, $k = 5$, $m = 17$, 95% CI [-0.15, 0.60]) and for self-efficacy ($g = 0.18$, $k = 3$, $m = 5$, 95% CI [-0.81, 0.91]). Most of the moderator variables were not significant. Importantly, to improve the quality of future reports on the effects of rubrics, we provide an instrument to be filled out for rubric scholars in forthcoming studies.

Keywords Rubrics · Meta-analysis · Academic performance · Self-regulated learning · Self-efficacy

✉ Leire Pinedo
leire.research@gmail.com

[1] Centre for Assessment Research Policy and Practice in Education (CARPE), School of Policy and Practice, Institute of Education, St. Patrick's Campus, Dublin City University, Dublin, Ireland

[2] Kristianstad University, Kristianstad, Sweden

[3] Facultad de Educación y Deporte, Universidad de Deusto, Bilbao, Spain

[4] Universidad Nacional de Educación a Distancia, Madrid, Spain

Introduction

Rubrics have become commonplace in educational settings across all instructional levels and throughout the world. Educators and policymakers put their trust in rubrics because rubrics are effective for judging the quality of student performance against pre-set criteria, increasing scoring reliability (Jonsson & Svingby, 2007). Additionally, the use of rubrics may also support students' academic achievement (e.g., Dawson, 2017).

Although there are several research reviews presenting evidence of positive effects on students' academic performance, self-regulated learning, and self-efficacy (e.g., Brookhart, 2018; Brookhart & Chen, 2015; Panadero & Jonsson, 2013; Reddy & Andrade, 2010), these reviews are all narrative, making it difficult to estimate the strength of this effect or the contribution from different moderating factors. Such estimations are especially important in a field where the findings are mixed, and the number of potential moderating factors is large, which is the case with research on the use of rubrics. Our aim is therefore to conduct a meta-analysis in order to estimate the strength of the effects on students' academic performance, self-regulated learning, and self-efficacy through the use of rubrics, while investigating the influence of moderating factors.

Definition of Rubrics and Terminology

Since it is not unusual that rubrics get confused with other tools such as checklist, rating scales, or performance list (Arter & McTighe, 2001; Brookhart, 2013, 2018), it is important to define what a rubric is. The definition of rubrics used in this study is taken from Brookhart (2018), where a rubric is described as a tool that: "articulates expectations for student work by listing criteria for the work and performance level descriptions across a continuum of quality" (p. 1). Rubrics are sometimes also described as containing "three essential features: evaluative criteria, quality definitions, and a scoring strategy" (Popham, 1997 p. 72). However, a rubric designed for formative purposes might not include an explicit scoring strategy, only assessment criteria, the corresponding performance levels, and a description on the intersections of criteria and performance levels. In this study, we will therefore not use the term "scoring rubrics," as this implies the existence of a scoring strategy.

The most commonly used representation of a rubric is a table or matrix. Usually, the first column to the left presents the assessment criteria. The rest of the columns contain the different performance levels, which can vary in their direction from high to low quality or, vice versa, low to high. The first row may contain labels for the performance levels (e.g., "outstanding," "deficient", etc.), and may also (in "scoring rubrics") include the points to be awarded for each level. Other, less common, designs also exist, which include the same features, but use a different representation (e.g., concentric circles).

Rubrics for formative purposes are usually created by the teachers, though other options are also possible, such as co-creation of rubrics with the students (e.g., Fraile et al., 2017) or implementing already existing rubrics from databases, such

as (for example) the Writing Rubrics for the English Language Proficiency Assessments for California.

The Origin and History of Rubrics

The term "rubric" has been found to be in use at least from the fifteenth century, referring to "headings of different sections of a book…Christian monks…initiating each major section of a copied book with a large red letter. Because the Latin word for red is ruber, rubric came to signify the headings for major divisions of a book." (Popham, 1997 p. 72). During the twentieth century, rubrics and similar tools were created to produce more reliable and valid inter-rater and intra-rater scores for qualitative judgments of student performance (i.e., a summative purpose). Turley and Gallagher (2008) refer to "A Scale for the Measurement of Quality in English Composition by Young People" by Hillegas (1912) as one of the first documented attempts to create a tool for increasing the objectivity in teachers' scoring. Another publication that has been pointed out as a key contribution for rubric development, is a study by Diederich and colleagues (1961), in which they summarized nearly a decade of research and conclude with five factors on judgments of writing ability: ideas, form, flavour, mechanics, and wording. According to Broad (2003), this was the inception of "standard, traditional, five-point rubric, by some version of which nearly every large-scale assessment of writing since 1961 has been strictly guided" (p. 6).

Prior to 1970, the term "rubric" only appeared in a few studies per year, but from thereon the term appears more and more frequently. From somewhere around the early years in the twenty-first century, the growth appears almost exponential and Dawson (2017) suggests that a seminal publication on rubrics by Popham (1997) may have sparked the rapid evolution of rubrics and the associated research during this time.

In his review, Dawson (2017) presented a figure on the cumulative frequency of articles and books including the term rubric. That figure contained information for articles until 2014 and for books until 2008. As a way to test whether the interest on rubrics is still increasing, we performed a similar search. The search method is explained in Appendix 1. As can be seen in Fig. 1, the use of the term "rubrics" in abstracts and as a topic, has been steadily growing over the years. Whether the decline after 2019 depends on the Covid pandemic, an actual decline in interest, or that the databases need some time to fully represent what has been published in the last couple years, or a combination of the previous, is not known.

In parallel to the increased interest in rubrics as a research topic, there has been an expansion of the use of rubrics across educational settings, from primary to higher education. As seen in the first research review on the use of rubrics (Jonsson & Svingby, 2007), this initial interest in rubrics was almost entirely for summative purposes. However, this was about to change, not least by the influential work by Grant Wiggins (1993, 1998), who on moral grounds argued against secrecy in assessment of student performance and advocated the use of rubrics to increase transparency in assessments, as well as to support formative feedback practices and

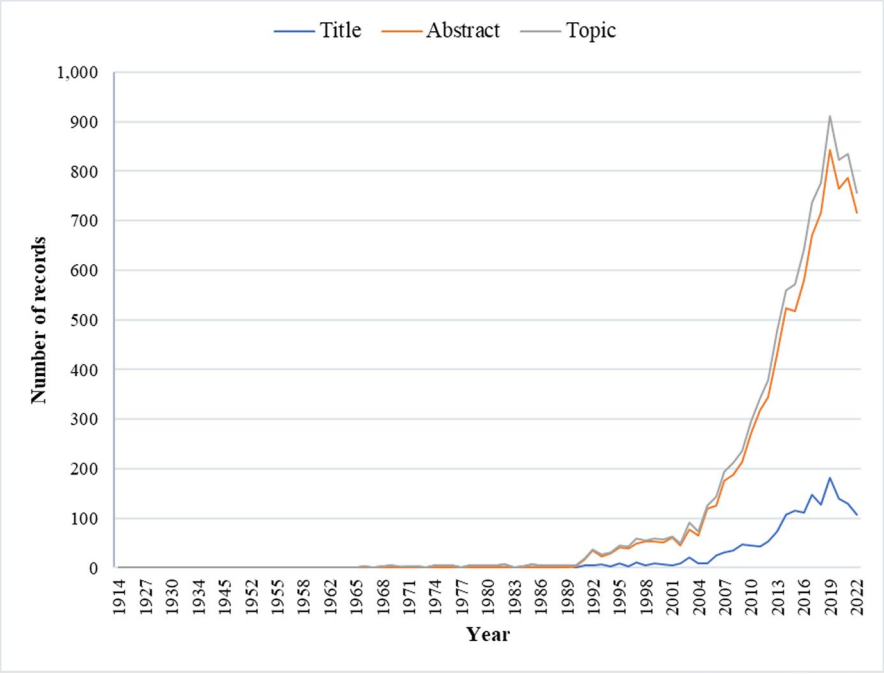

Fig. 1 Publications per year

student self-regulation. This growing interest in the formative use of rubrics can be seen in the review mentioned above, where Jonsson and Svingby (2007) found merely 15 studies with this focus, while a decade later Brookhart (2018) found 46, even though she searched for studies in a higher education context only.

The increased use of rubrics for formative purposes has also sparked a heated debate regarding the merits of transparency and use of rubrics to support student performance, where some authors claim, for instance, that sharing criteria with students is not beneficial and almost inevitably leads to instrumental learning and "criteria compliance" among students (e.g., Sadler, 2014; Torrance, 2007). However, by reviewing the arguments against the use of rubrics, Panadero and Jonsson (2020) show that the empirical evidence behind this critique is currently weak, and mainly based on anecdotal evidence and/or personal experiences. This debate is therefore likely to intensify, especially if the use of rubrics for formative purposes continues to grow.

Effects of Rubrics on Student Performance

In general, it has been argued that the bigger effects of rubrics on students' learning occur when they are implemented for self-assessment or peer assessment purposes. In their review of research on the formative use of rubrics, Panadero and Jonsson (2013) suggest—in line with the arguments made by Wiggins (1998)—that rubrics

provide a measure of transparency to the assessment situation. By making criteria, performance levels, and (when relevant) scoring strategies explicit, these may become objects of action and reflection themselves (i.e., students can use them to regulate their learning) helping students to improve their learning via self and peer assessment (Nicol, 2021; Panadero et al., 2019). This interpretation is supported by students, who are generally positive about being provided with rubrics and claim to use the rubrics to better understand (and meet) expectations (e.g., Andrade & Du, 2005; Jonsson, 2014; Reynolds-Keefer, 2010). Consequently, there is a number of studies where rubrics have been used to promote different aspects of self-regulated learning, such as establishing more accurate goals, monitoring progress, and increasing self-regulation and metacognition (Brookhart & Chen, 2015; Panadero & Jonsson, 2013), while reducing cognitive load (Krebs et al., 2022), and increasing self-efficacy (Reddy & Andrade, 2010). All of these aspects are likely to contribute to improved performance among students (Panadero & Jonsson, 2013).

Several rubric studies have found a positive impact on student performance, where the students using a rubric showed a greater improvement compared to a control group (e.g., Brookhart & Chen, 2015). However, the findings from other studies are not necessarily as straight forward, for instance by only reporting statistically significant effects for some criteria or some groups of students (e.g., Becker, 2016; Montanero et al., 2014; Sáiz-Manzanares et al., 2015). Furthermore, longer and larger interventions might be needed in order to produce positive results for younger students (Panadero & Jonsson, 2013).

In addition to encompassing many different educational levels, topics, research designs, etc., most studies combine the use of rubrics with other pedagogical tools, such as "model assignments" (Andrade et al., 2008), feedback (Panadero et al., 2012), or "Jonassen's (1999) constructivist, reflective, purposeful, interactive cooperative learning environment model" (Bay & Kotaman, 2011 p. 286). This diversity makes it difficult to get an overview of the effects on student performance by the use of rubrics from the narrative reviews currently available, or the influence from different moderating factors.

In addition, since the transparency provided by rubrics is thought to support students' self-regulation of learning (SRL) and self-efficacy (Panadero & Jonsson, 2013), thereby indirectly affecting students' academic performance, both variables are of particular importance in this context. As mentioned, rubrics have been used to promote self- and peer assessment, as rubrics may help students to establish more accurate goals, monitor their progress, increase self-regulation and metacognition (Brookhart & Chen, 2015; Panadero & Jonsson, 2013). This points to the importance of considering self-, and peer assessment as moderators, as well as SRL-variables and self-efficacy as important outcomes of using rubrics, along with student performance.

Potential Moderator Variables

Given the extensive use of rubrics, a great variation in design and implementation is to be expected, including different moderating factors that may have an impact on

the effects of using rubrics. For this reason, we explore eleven variables as possible moderators on the effect of using rubrics on academic performance. First, the year of publication could influence the effect found in rubric research. As shown by Dawson (2017), rubric research has become more and more widespread, which means that, by "standing on the shoulders of previous research", the quality of interventions may have improved over time. Therefore, we will explore whether interventions are getting more precise and stronger over time by exploring the year of publication as a moderator.

Second, some studies have found different effects for boys and girls, where girls benefit more from the intervention (e.g., Andrade et al., 2009). A similar effect is reported in a meta-analysis on the effects of self-assessment interventions (Panadero et al., 2017). Some peer assessment studies have also found gender differences (Bloom & Hautaluoma, 1987; Wen & Tsai, 2006). As rubrics are often implemented in combination with self and peer assessment activities, we included gender as a moderating factor.

Third, mean age needs to be considered, as research seems to suggest that students in more advanced educational levels may benefit more from self- and peer assessment interventions (Yan et al., 2022). As the use of rubrics facilitates different aspects of self-regulation, it is of interest to explore if a similar pattern can be identified in relation to the use of rubrics. Fourth, educational level, obviously in close relationship with mean age, will also be explored as a potential moderator. Fifth, an additional aspect related to age, but also to educational level, is whether the rubric intervention takes place in compulsory education (e.g., primary, secondary) or higher education. To our knowledge, there are no studies on rubrics that empirically compare these levels, but using a meta-analytical approach, we can explore such relationships.

Sixth, the number of sessions in which the students use the rubric is another moderating factor that could influence the effect. Just handing out rubrics does not guarantee educational gains (Brookhart, 2018). Therefore, it could be expected that students need to work with the rubrics for a number of sessions before an effect can be documented.

Seventh and eighth, rubrics can vary in relation to the number of assessment criteria and performance levels (Brookhart, 2013). There are only general indications on what could be the adequate number, and this could be greatly influenced by the context and students' knowledge. Therefore, we investigate if the number of criteria or performance levels make a difference in the effect of using rubrics.

Ninth, when implemented with the intention of improving students' learning, rubrics are supposed to generate positive effects in combination with students' self and peer assessment (Brookhart, 2018). As these two are different in nature and may affect the outcome, we will explore them as potential moderators.

Tenth, we also explore research design as a potential moderator. Stronger research designs (i.e., randomized controlled trials) have more control over potential cofounding variables and may therefore lead to more accurate results. Thus, we explore if there are differences based on research design.

Finally, eleventh, the empirical quality of the study can influence the effects of an intervention (Valentine, 2020). More robust research designs with a larger sample,

more acute operationalizations of the variables, more precise measurements, etc., might be more capable of identifying educational effects with greater precision. The instrument designed and used to measure empirical quality can be found in Appendix 2.

Another moderator that we considered was the timing of the rubric. Studies can be divided into two broad categories: (1) studies where the students are provided with rubrics before task performance, and therefore can use the rubric to plan and monitor their performance; and (2) studies where the students use rubrics to revise (and hopefully improve) previous task performance (Jonsson, 2020). Unfortunately, when we coded this moderator, we found that in the vast majority of studies the rubric was provided before, making it statistically unfeasible to run the analysis.

Aim and Research Questions

Our aim is to investigate the effects of using rubrics on students' academic performance, self-regulated learning, and self-efficacy through three meta-analyses. We formulated three research questions (RQ):

RQ1. What is the effect of using rubrics and its potential moderators on academic performance?
RQ2. What is the effect of using rubrics and its potential moderators on self-regulated learning?
RQ3. What is the effect of using rubrics on self-efficacy?

Method

Selection of Studies

The search was conducted in February 2022 in PsycINFO, ERIC, SCOPUS, Web of Science, and ProQuest databases. Authors used the following combinations of keywords for the fields of title and abstract: rubric* OR (assessment AND matrix) OR (scoring AND guide*) OR (scoring AND grid) OR (grading AND list). Additionally, since 2022 the first and second authors have been particularly attentive to any new publication on rubrics via journal and citation alerts.

The inclusion criteria were: (a) the study included empirical results on the use of rubrics in relation to cognitive and non-cognitive skills, (b) analyzed the individual use of rubrics (not group work), (c) had a control group, (d) had quasi- or experimental designs, (e) had been peer-reviewed, and (f) was published in English.

Figure 2 is a flowchart of the search and inclusion process. In total, 3,848 records were identified from the search. After removing duplicates, the first two authors independently screened the abstract of 200 out of 2,564 papers to assess the eligibility of the selected papers until a 100 percent agreement was reached. The remaining papers were divided among the first two authors to select the relevant empirical studies. The full texts of 97 selected studies were read and assessed. Reference

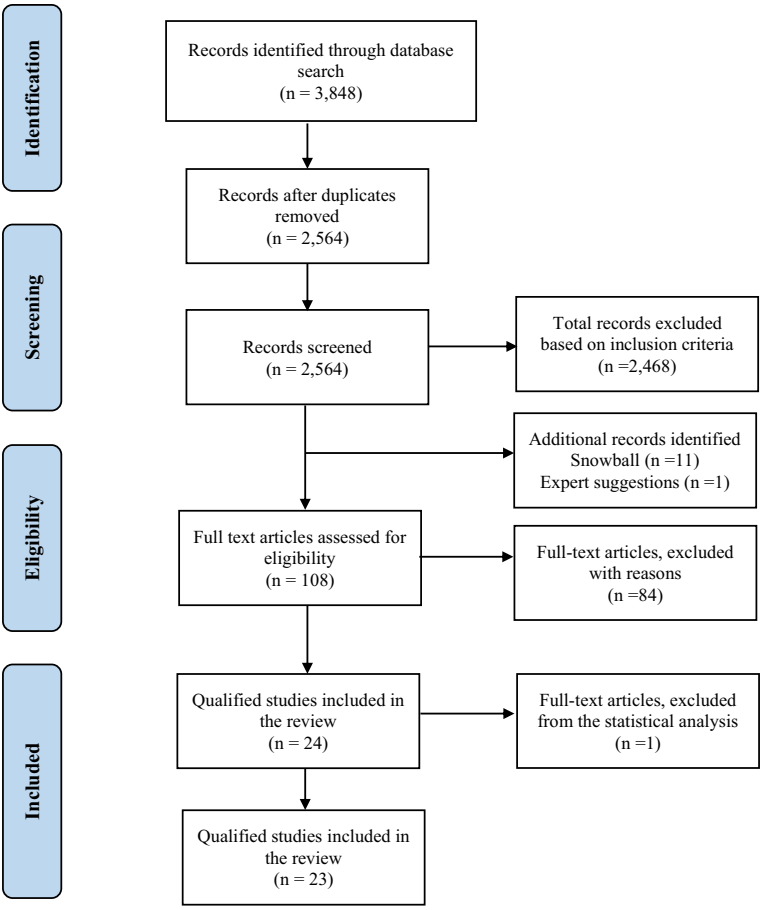

Fig. 2 Flowchart of search process and inclusion of articles

lists of empirical articles and quotes from the included articles were also examined, resulting in 12 publications of interest. Based on the inclusion criteria, 24 studies were included in the data set. However, only 23 studies were included in the meta-analyses as there was only one study (Hongxia et al., 2020) exploring "engagement" as a variable, so it was not feasible to explore this variable due to the low number of studies. Of these 23 studies, 21 provided effect sizes for academic performance (54 effect sizes), 5 reported results on self-regulated learning (17 effect sizes), and 4 studies (with one effect size per study) reported results on self-efficacy.

If the studies did not include enough information to calculate an effect size, we contacted the authors and asked for the information. A total of thirteen authors were contacted, eight for statistical information, and five for information about the empirical procedure. Eleven authors replied, but only six provided the necessary information.

Coding the Articles

The first and second authors independently assessed the eligibility of 200 papers in two rounds, reaching a 98.5% inter-rater agreement. The screening of the remaining records was equally distributed between the first and second authors who also coded all the selected articles using a database containing the following categories: (a) aim; (b) research questions; (c) hypothesis; (d) characteristics of the sample (country of origin, percentage of women in the sample, mean age, grade, educational level); (e) self-assessment and peer assessment terminology, use of peer-assessment and whether they included grades; (f) use of other aids (exemplars, innovative computer scaffold, etc.); (g) length of intervention/training measured in number of sessions; (h) task and subject; (i) independent and dependent variables; (j) type and time of measurement, and type of data collection; (k) research design (type, use of pre-test and post-test and longitudinal data); (l) whether gender differences were found; (m) observations on the procedure; (o) results; (p) rubric characteristics (time when it was handed out, number of assessment criteria and levels); (q) conclusions; (r) direction of effect (whether dependent variables were inversed), (s) quality of the studies according to an "empirical quality" rubric (Panadero et al., 2023) and slightly modified for this study; and (t) statistical data (type, mean score, standard deviation, sample size, and Pearson correlation). A series of meetings among all the authors were held to discuss relevant aspects of the coding to maintain rigor throughout the process.

From the aforementioned variables, the following characteristics of the studies were further operationalized to be used as potential moderator variables: percentage of females in the sample, mean age of the sample, mean age imputed (i.e., when age was not reported, it was inferred from the specific educational grade[1]), educational level with the following categories: primary (from grade 1 to 6), middle (from grade 7 to 8), secondary (from grade 9 to 12), or higher education,[2] type of educational level (compulsory vs. higher education), length of the intervention (measured in number of sessions, and when it was not reported it was inferred from the intervention duration[3]), performance levels of the rubric (the levels of executions usually presented as the number of columns), number of assessment criteria in the rubric (the areas to be evaluated usually presented in the first column), research design (quasi-experimental or experimental), and an empirical quality score obtained from an instrument designed for this study (Appendix 2).

This "empirical quality" instrument consists of a rubric with the following categories: research design, quality of the measurement tools, quality of the intervention (three subcategories: training, practice, length of the intervention), quality of

[1] Specifically, a mean age of 9.5, 13.5, and 15.5 was imputed for grades four, eight, and tenth, respectively.

[2] Two studies were excluded from this moderator analysis because they included students from more than one educational level (primary & secondary, and primary & middle).

[3] An intervention duration between 40 min to 1 h was coded as one session. Seven studies have 1 session, four studies between 2 and 3 sessions, and seven studies more than 4 sessions.

dependent variables, and quality of the sample. Each study was evaluated according to all of these categories to calculate a summary score for quality.

The data and coding of all studies, including the measures of quality, can be found in the following link: https://osf.io/d6mry/?view_only=0db0f9c5617a492c84500c0a83a185b9.

Statistical analyses

We calculated and synthesized the standardized mean difference (Cohen's *d*) as the main effect size, which indicates the magnitude of the difference in the dependent variable between the control group and the intervention group at a specific point in time. We applied Hedges' correction to avoid overestimated standardized mean differences (Hedges, 1981), and the effect sizes are therefore referred to as Hedges' *g*. A positive Hedges' *g* indicates that the performance (or other outcome) of the intervention group was higher as compared to the control group. In addition, we calculated the standardized mean change (Becker, 1988)[4] for studies reporting pre-post measures for both the control and the intervention group. A standardized mean change indicates how much the outcome variable of the intervention group improved with respect to the improvement observed in the control group. A positive standardized mean change indicates that the outcome of the intervention group has improved to a greater extent (from the pre- to the post-test) as compared to the control group. To calculate the sampling variance of the standardized mean change, information on the correlation between pre-post measures in each group is needed. Since this correlation was not reported in the primary studies, we imputed a correlation of 0.4, which represents a moderate-to-high correlation according to Cohen's cutoffs (Cohen, 1988). To interpret the overall effect size, we used the empirical cutoffs proposed by Hattie (2012). Specifically, Cohen's *d* values from 0 to 0.2 were considered low, from 0.2 to 0.4 medium, and above 0.4 high (or "desired") overall effect. We report overall effects together with their 95% confidence intervals and 95% prediction intervals. Prediction intervals indicate the range in which the effect size of a future study would be expected to fall within, if this study was chosen from the same population of studies included in one of these meta-analyses.

In some studies, several effect sizes could be calculated, either because the dependent variable was measured using different indicators, or because different types of rubrics were used (leading to multiple comparisons within that study). To model the dependency among effect sizes that were extracted from the same study, we applied a meta-analytic three-level model to synthesize effect sizes (Cheung, 2014; Van den Noortgate et al., 2013, 2015). This model takes into account the variability between effect sizes within studies (Level 2), and the variability among study effects across studies (Level 3). To test whether the variance within and between studies were significantly different from zero, we applied likelihood ratio tests,

[4] To calculate a standardized mean change, you first need to calculate the difference between the pre- and post-measures within groups. To calculate these standardized changes within groups, the standard deviation of the pre-test as the denominator has been used, as suggested by Becker (1988).

comparing the fit of a model that considers both sources of variance with the fit of a model that ignores one of these variances. If substantial variability among effect sizes was detected, we entered the moderator variables one by one in the three-level model. Continuous variables were previously centered. To avoid type I error rates above the nominal level ($\alpha = 0.05$), we applied the robust variance correction a posteriori to all these analyses (Fernández-Castilla et al., 2021a, 2021b; Tipton et al., 2019). Therefore, all the standard errors and confidence intervals reported in the result section are those obtained after applying the robust variance correction. At least ten effect sizes had to be available for each meta-regression to ensure minimal statistical power to detect an effect.

An analysis of outliers was carried out to detect extreme effect sizes. Specifically, we calculated the studentized deleted residuals for each effect size (Viechtbauer & Cheung, 2010), and we searched for effect sizes with studentized deleted residuals beyond ± 1.96. If potential outliers were detected, we removed them, and repeated all analyses without them.

Finally, we implemented several procedures to check for the presence of publication bias. First, we visually evaluated a figure known as the funnel plot, where effect sizes are plotted against their corresponding standard errors. In cases without publication bias, the distribution of effect sizes should exhibit symmetry within the funnel plot. When the distribution of effect sizes is asymmetrical—indicating a higher number of small studies contributing significant positive effects compared to small studies with considerable negative effects—it raises the possibility of the presence of publication bias. Second, we calculated the L_0^+ statistic of the Trim and Fill method (Duval & Tweedie, 2000), as according to Fernández-Castilla et al. (2021a, 2021b) this is currently the method with the highest power to detect the presence of publication bias (once controlled for the Type I error) given the characteristics of our dataset. If the value of L_0^+ was 3 or larger, we concluded that publication bias could exist. If publication bias was detected, we calculated a corrected overall effect size using the selection method of Vevea and Woods (2005). More information about how these methods can be used to detect and correct for publication bias can be found in the Supplementary material (Appendix 3).

All the analyses were carried out in R studio, using the following packages: *metafor* (Viechtbauer, 2010), *clubSandwich* (Pustejovsky, 2021), and *weightr* (Corbun & Vevea, 2019).

Results

RQ1. What Is the Effect of Using Rubrics and Its Potential Moderators on Academic Performance?

Meta-analysis on Standardized Mean Differences A total of 54 standardized mean differences (Hedges' g) were synthesized from 21 studies. The study-effects are depicted in Fig. 3. The black confidence interval around each study-effect represents the global study precision, whereas the gray confidence interval represents the

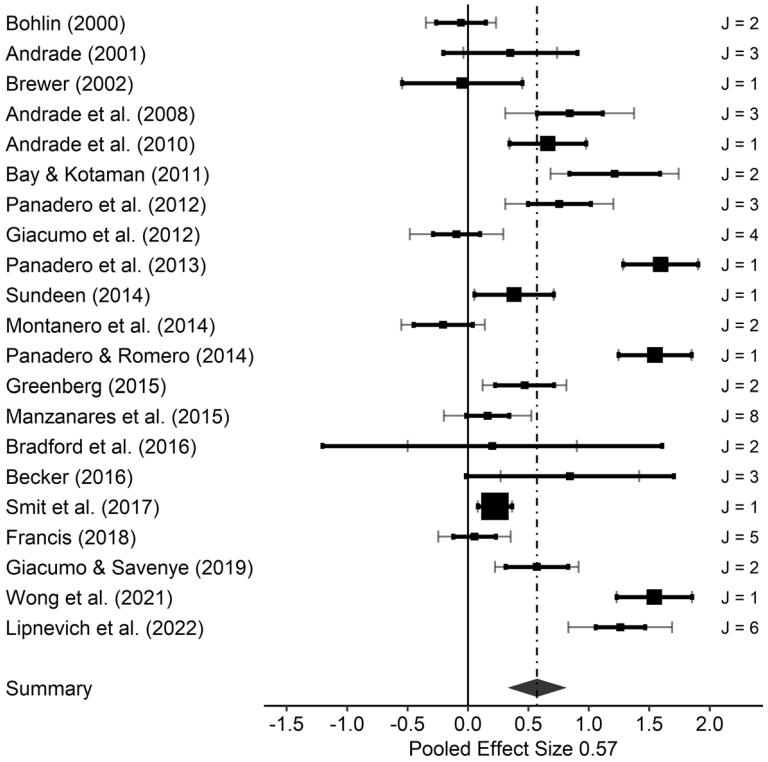

Fig. 3 Forest plot of studies reporting effects on academic performance (standardized mean differences). Note: Black confidence intervals represent the global study-precision, whereas gray confidence intervals have been calculated using only sample size information. Therefore, the width of the gray confidence intervals is a direct function of the sample size, whereas the width of the black confidence intervals is a direct function of the global precision of the study, which depends on the sample size, the within study variability in that study, and the number of effect sizes reported. The dimensions of the squares correspond to the weights assigned to each study when computing the pooled effect size (larger squares indicate higher assigned weight). The J index denotes the count of effect sizes reported within the studies

precision provided only by the sample size of each study (Fernández-Castilla et al., 2020).

The overall Hedges' g was 0.57 (SE = 0.125, $t = 4.59$, p-value < 0.001, 95% CI [0.312, 0.831], 95% PI [-0.558, 1.701]), which is a large overall effect. The variability between study-effects was statistically different from zero ($\widehat{\sigma}^2_{between} = 0.289$, LRT = 4.22, p-value < 0.05) and the within-study variability as well ($\widehat{\sigma}^2_{within} = 0.028$, LRT = 29.15, p-value < 0.001). No outliers were detected, and none of the potential moderator variables were significantly related to the observed effect sizes (see Table 1).

A slight asymmetry was detected in the visual inspection of the funnel plot (see Fig. 4): smaller studies tend to have larger effect sizes. However, the L_0^+

Table 1 Results from the moderator analyses for academic performance

	M	*g* (SE)	*t* (*p* -value)	95% CI	$\widehat{\sigma}^2_{between}$	$\widehat{\sigma}^2_{within}$
Publication year					0.266 (21)	0.027 (54)
Intercept	54	0.598 (0.123)	4.86 (<0.001)	[0.34, 0.86]		
Year[c]		0.033 (0.014)	2.35 (0.060)	[-0.01, 0.07]		
Gender					0.171 (16)	0.054 (40)
Intercept	40	0.558 (0.120)	4.42 (<0.001)	[0.29, 0.83]		
Perc. Females[c]		0.008 (0.004)	1.06 (0.420)	[-0.03, 0.05]		
Mean age					0.401 (5)	0.026 (19)
Intercept	19	1.023 (0.333)	3.07 (0.060)	[-0.09, 2.13]		
Mean age[c]		-0.042 (0.072)	-0.59 (0.617)	[-0.34, 0.26]		
Mean age (imputed)					0.249 (16)	0.045 (40)
Intercept	40	0.584 (0.147)	3.98 (0.002)	[0.26, 0.91]		
Mean age imp.[c]		0.019 (0.019)	1.03 (0.378)	[-0.04, 0.08]		
Educational level					0.277 (20)	0.026 (53)
Primary	10	0.196 (0.185)	1.06 (0.350)	[-0.32, 0.71]		
Middle	10	0.819 (0.256)	3.20 (0.086)	[-0.29, 1.92]		
Secondary	4	0.127 (0.182)	0.70 (0.612)	[-2.18, 2.43]		
Higher	29	0.756 (0.206)	3.68 (0.005)	[0.29, 1.22]		
There are no statistical differences between categories						
Type educational level					0.294 (20)	0.026 (53)
Compulsory	24	0.374 (0.152)	2.46 (0.036)	[0.03, 0.72]		
Higher ed	29	0.756 (0.206)	3.68 (0.005)	[0.29, 1.22]		
Number sessions					0.206 (17)	0.033 (48)
Intercept	48	0.791 (0.205)	3.86 (0.002)	[0.34, 1.24]		
Sessions[c]		-0.059 (0.024)	-2.44 (0.059)	[-0.12, 0.01]		
Performance levels					0.323 (21)	0.027 (54)
4	35	0.593 (0.153)	3.88 (0.002)	[0.27, 0.92]		
5	18	0.537 (0.267)	2.01 (0.138)	[-0.32, 1.39]		
6	1	0.381	-	-		
Assessment criteria					0.274 (19)	0.032 (51)
Intercept	51	0.609 (0.127)	4.81 (<0.001)	[0.34, 0.88]		
Assessment[c]		0.036 (0.015)	2.50 (0.157)	[-0.04, 0.11]		
Use of SA, PA or both					0.271 (21)	0.027 (54)
Self-assess	46	0.288 (0.510)	0.566 (0.672)	[-6.19, 6.76]		
Peer-assess	5	0.521 (0.122)	4.269 (<0.001)	[0.26, 0.78]		
Both	3	1.072 (0.513)	2.090 (0.172)	[-1.14, 3.29]		
Research design					0.297 (21)	0.028 (54)
Experimental	17	0.705 (0.228)	3.08 (0.027)	[0.12, 1.29]		
Quasi-experi	37	0.518 (0.151)	3.43 (0.004)	[0.19, 0.84]		
Empirical quality					0.285 (21)	0.027 (54)
Intercept	54	0.561 (0.123)	4.57 (<0.001)	[0.30, 0.82]		
Quality[c]		-0.060 (0.050)	-1.21 (0.254)	[-0.17, 0.05]		

Upper script *c* means that the variable has been centered prior the analyses. *M* = number of effect sizes involved in each analysis. $\widehat{\sigma}^2_{within}$ = within-study variance; $\widehat{\sigma}^2_{between}$ = between-studies variance.

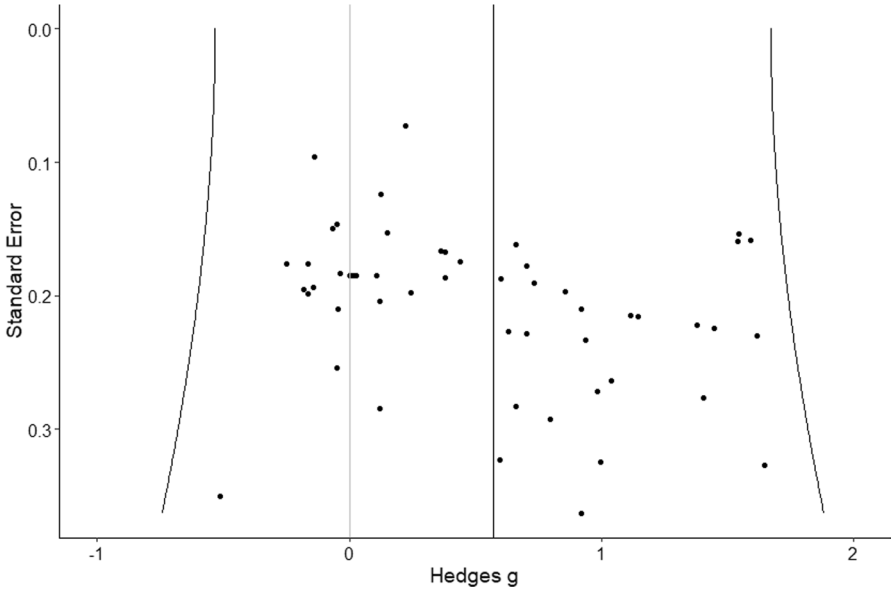

Fig. 4 Funnel plot of standardized mean differences (Hedges' *g*) for academic performance

statistic was 0, meaning that the Trim and Fill test did not detect the presence of publication bias. Given the mixed evidence on the potential presence of publication bias, we decided to apply Vevea and Woods (2005) selection model to obtain a corrected, overall effect size. The corrected overall effect was 0.45 for moderate bias, and 0.33 for severe bias.

Meta-analysis on Standardized Mean Changes A second meta-analysis was calculated with the twelve studies including 30 effect sizes that reported pre- and post-measures from both the intervention and control group, as in these studies standardized mean changes could be obtained. Figure 5 illustrates the study-effects of this subset of studies.

The overall standardized mean change was 0.38 (SE = 0.16, $t = 2.34$, p-value = 0.042, 95% CI [0.02, 0.75], 95% PI [-0.75, 1.51]), which is a medium overall effect. The variability between study-effects was statistically different from zero ($\hat{\sigma}^2_{between} = 0.207$, LRT = 6.01, p-value = 0.014), as well as the within-study variability ($\hat{\sigma}^2_{within} = 0.098$, LRT = 9.06, p-value = 0.003). Once again, no outliers were detected.

The mean age of the participants significantly predicted the variability of the standardized mean changes: for a one unit increase in the average age, there was a 0.062 unit decrease in the standardized mean changes, which suggests that the effect of the use of rubrics on academic performance decreases as the mean age of the sample increases (SE = 0.001, $t = -243$, $p = 0.001$). However, these results must be interpreted with caution, since they are based on only 14 effect

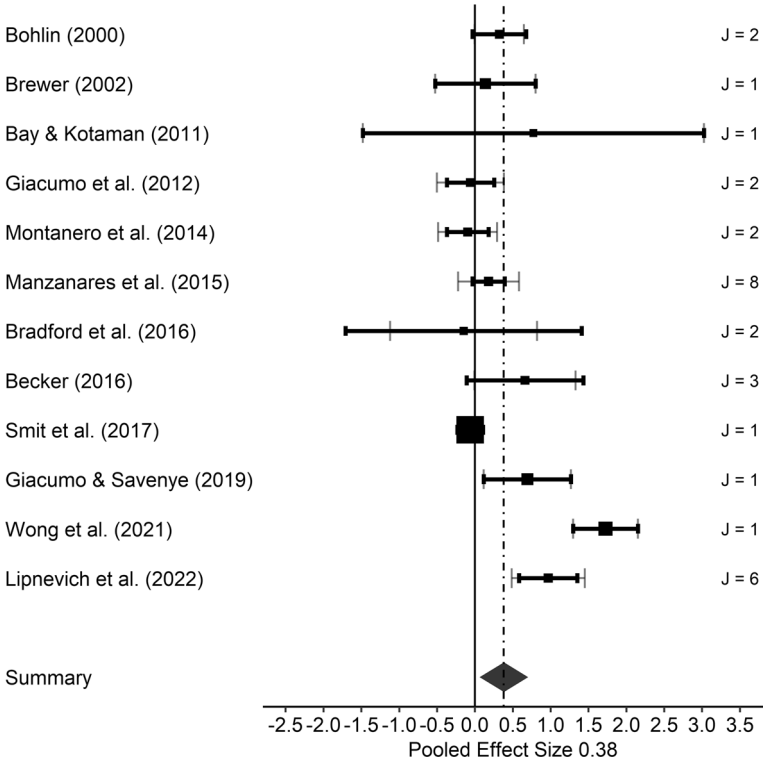

Fig. 5 Forest plot of studies reporting effects on academic performance for which standardized mean changes could be calculated. Note. Black confidence intervals represent the global study-precision, whereas gray confidence intervals have been calculated using only sample size information. Therefore, the width of the gray confidence intervals is a direct function of the sample size, whereas the width of the black confidence intervals is a direct function of the global precision of the study, which depends on the sample size, the within study variability in that study, and the number of effect sizes reported. The dimensions of the squares correspond to the weights assigned to each study when computing the pooled effect size (larger squares indicate higher assigned weight). The J index denotes the count of effect sizes reported within the studies

sizes from two studies, and there were only two different values for the mean age. The overall standardized mean change also differed across educational levels: the overall effect for studies from secondary education was higher ($g = 0.99$, number of effects = 10) than the overall effect of studies from primary education ($g = 0.03$, number of effects = 10, difference = 0.96, $p = 0.004$) and middle school ($g = 0.14$, number of effects = 4, difference = 0.85, $p < 0.001$). No other effects were observed for the rest of the study variables (the results are provided in the supplementary material Appendix 3).

The funnel plot (Fig. 6) seems symmetric, meaning that publication bias might not be present. Furthermore, the Trim and Fill — L_0^+ statistic is equal to 0, which is another indicator suggesting that there is no publication bias.

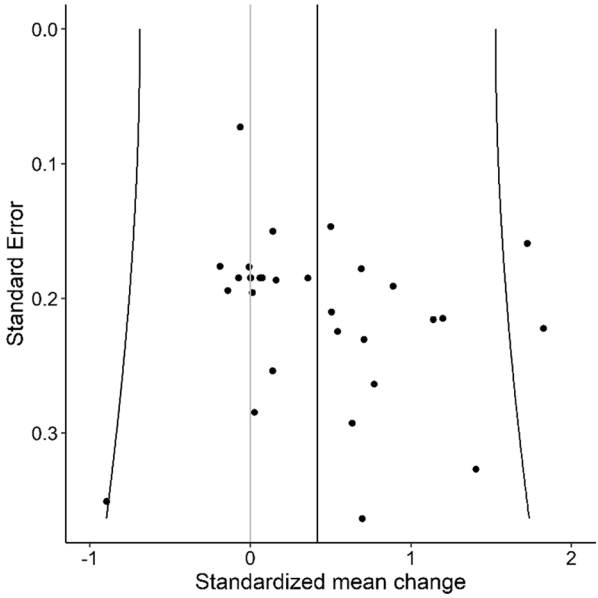

Fig. 6 Funnel plot of standardized mean changes for academic performance

RQ2. What Is the Effect of Using Rubrics and Its Potential Moderators on Self-Regulated Learning?

A total of 17 effect sizes from five studies were synthesized. The study-effects are depicted in Fig. 7. The overall standardized mean difference was 0.23 (SE = 0.123, $t = 1.82$, p-value = 0.156, 95% CI [-0.15, 0.60], 95% PI [-0.65, 1.10]), which is small in magnitude. The variability between study-effects was not significantly different from zero ($\hat{\sigma}^2_{between} = 0.017$, LRT = 6.01, $p = 0.014$), but the within-study variability was ($\hat{\sigma}^2_{within} = 0.168$, LRT = 73.25, $p < 0.001$). No outliers were detected.

Only one moderator variable explained part of the observed variability: when self-assessment and peer-assessment were used, the overall effect was significantly larger than when only self-assessment was used (see Table 2). However, since there are only five effect sizes available for the category "self-assessment and peer-assessment used", these results should be interpreted with caution. The effect sizes seem to symmetrically distributed across the funnel plot (Fig. 8), meaning that publication bias is not likely not be present. This result is further supported by the Trim and Fill — L_0^+ statistic which is equal to 0.

Meta-analysis on Standardized Mean Changes Only three studies provided enough information to calculate a standardized mean change (including five effect sizes). These five effect sizes were pooled using a standard random effects model. The pooled effect was 0.05 (SE = 0.189, Z = 0.263, $p = 0.792$, 95% CI [-0.32, 0.42], 95% PI [-0.81, 0.91], $\hat{\sigma}^2_{between} = 0.156$), which is a small (almost null) effect. Since the number of effect sizes available was very small, meta-regressions and publication bias analyses were not carried out.

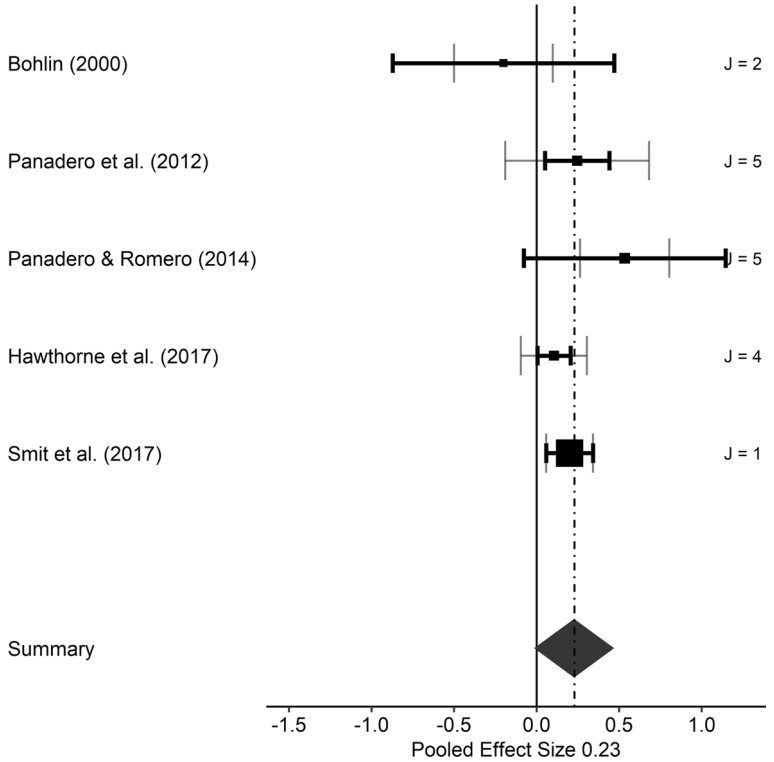

Fig. 7 Forest plot of studies reporting effects on self-regulated learning (standardized mean differences). Note. Black confidence intervals represent the global study-precision, whereas gray confidence intervals have been calculated using only sample size information. Therefore, the width of the gray confidence intervals is a direct function of the sample size, whereas the width of the black confidence intervals is a direct function of the global precision of the study, which depends on the sample size, the within study variability in that study, and the number of effect sizes reported. The dimensions of the squares correspond to the weights assigned to each study when computing the pooled effect size (larger squares indicate higher assigned weight). The J index denotes the count of effect sizes reported within the studies

RQ3. What Is the Effect of Using Rubrics on Self-Efficacy?

Only four effect sizes (coming from four different studies) summarized the effect of using rubrics on self-efficacy (see Fig. 9). Therefore, a standard random-effects model was applied. The overall effect was 0.18 (SE = 0.05, Z = 3.31, $p < 0.001$, 95% CI [0.07, 0.29], 95% PI [0.07, 0.29], $\hat{\sigma}^2_{between} = 0.000$), which is small in magnitude. No other analyses were performed due to the small number of effect sizes.

Secondary Results from* a Posteriori *Hypothesis Moderator analyses in relation to standardized mean changes of academic performance suggested that the effect of using rubrics decreases as the mean age of the sample increases. However, the overall effect for studies from secondary education was also higher than those from primary education, making the results contradictory. As it has been suggested in

Table 2 Results from the moderator analyses for self-regulated learning

	M	*g* (SE)	*t* (*p*-value)	95% CI	$\widehat{\sigma}^2_{between}$	$\widehat{\sigma}^2_{within}$
Publication year					0.166 (5)	0.018 (17)
Intercept	17	0.228 (0.129)	1.77 (0.190)	[-0.22, 0.68]		
Year[c]		0.025 (0.024)	1.01 (0.457)	[-0.14, 0.19]		
Gender					0.149 (4)	0.000 (15)
Intercept	15	0.298 (0.054)	5.57 (0.031)	[0.07, 0.53]		
Perc. females[c]		0.013 (0.004)	2.92 (0.157)	[-0.02, 0.04]		
Mean age					0.050 (2)	0.247 (10)
Intercept	10	0.389 (0.002)	236.3 (0.001)	[-0.09, 2.13]		
Mean age[c]		0.046 (0.001)	86.9 (0.004)	[-0.34, 0.26]		
Mean age (imp)					0.000 (4)	0.204 (13)
Intercept	13	0.285 (0.038)	7.46 (0.042)	[0.03, 0.54]		
Mean age imp.[c]		0.049 (0.011)	4.39 (0.063)	[-0.01, 0.11]		
Educational level					0.033 (5)	0.170 (17)
Primary	3	-0.047 (0.283)	-0.24 (0.849)	[-2.53, 2.43]		
Middle	5	0.246 (0.001)	729.4 (<0.001)	[0.24, 0.25]		
Secondary	0	-	-	-		
Higher	9	0.328 (0.214)	1.53 (0.368)	[-2.39, 3.05]		
There are no statistical differences between categories						
Type educational level					0.025 (5)	0.166 (17)
Compulsory	8	0.104 (0.153)	0.69 (0.575)	[-0.69, 0.90]		
Higher ed	9	0.329 (0.214)	1.53 (0.368)	[-2.39, 3.05]		
Number sessions					0.000 (3)	0.239 (11)
Intercept	11	0.422 (0.161)	2.62 (0.232)	[-1.62, 2.47]		
Sessions[c]		-0.025 (0.018)	-1.38 (0.400)	[-0.25, 0.20]		
Performance levels					0.031 (5)	0.170 (17)
4	13	0.256 (0.158)	1.62 (0.220)	[-0.31, 0.82]		
5	4	0.104 (0.001)	632.1(0.001)	[0.10, 0.11]		
6	0	-	-	-		
Assessment criteria					0.000 (4)	0.145 (15)
Intercept	15	0.269 (0.040)	7.4 (0.038)	[0.05, 0.54]		
Assessment[c]		0.354 (0.060)	5.9 (0.052)	[-0.01, 0.72]		
Use of SA, PA or both					0.000 (5)	0.153 (17)
Self-assessment	12	0.114 (0.086)	1.33 (0.302)	[-0.22, 0.44]		
Peer-assessment	0	-	-	-		
Both	5	0.532 (0.003)	202.95 (0.003)	[0.50, 0.57]		
Research design					0.033 (5)	0.171 (17)
Experimental	9	0.175 (0.071)	2.46 (0.246)	[-0.73, 1.08]		
Quasi-experi	8	0.259 (0.264)	0.98 (0.446)	[-1.11, 1.63]		
		There are no statistical differences between categories				
Empirical quality					0.000 (5)	0.158 (17)
Intercept	17	0.244 (0.070)	3.49 (0.052)	[-0.01, 0.49]		
Quality[c]		-0.080 (0.044)	-1.82 (0.207)	[-0.26, 0.10]		

Upper script *c* means that the variable has been centered prior the analyses. $\widehat{\sigma}^2_{within}$ = within-study variance; $\widehat{\sigma}^2_{between}$ = between-studies variance, SA = self-assessment, PA = peer-assessment.

Significant differences between categories (difference = 0.418, SE = 0.086, $t = 4.88$, $p = 0.030$)

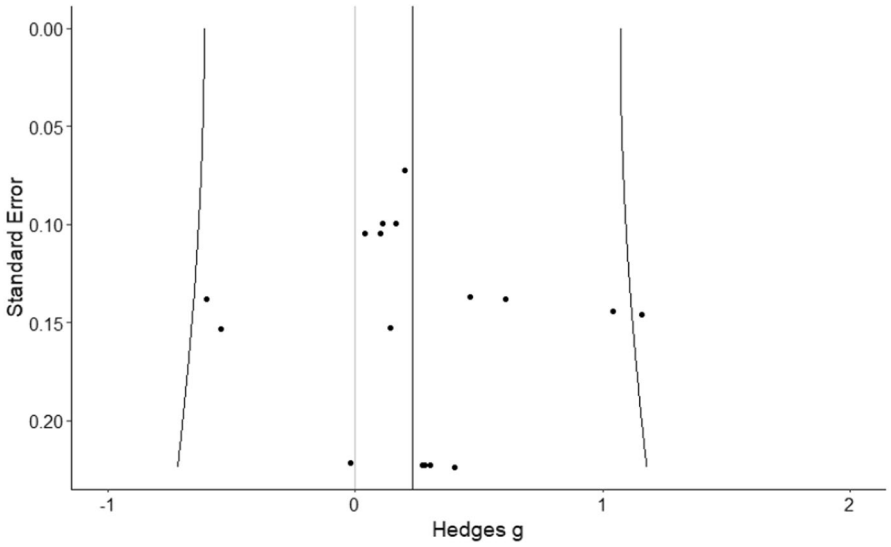

Fig. 8 Funnel plot of standardized mean differences (Hedges *g*) for self-regulated learning

previous research (Panadero & Jonsson, 2013) that there is an interaction between age/educational level and the "time devoted to work with the rubric" (p. 140), secondary analyses were carried out, using the standardized mean differences (Hedges' *g*) for academic performance, to explore this interaction.

The number of sessions (centered), mean age (centered), and the interaction were therefore entered in the meta-regression, producing a statistically significant and negative interaction effect (B = -0.041, SE = 0.001, t = -62.1, $p < 0.001$, 95% CI [-0.05, -0.03]), which means that the relationship between the number of sessions and the standardized mean differences decreases as the mean age increases (see Fig. 10). That is, the length of the intervention has less effect on the observed effect in samples with a larger mean age. Interpreting it in another way, this negative interaction effect indicates that the effect of the mean age on the observed effect sizes increases as the length of the intervention diminishes. When mean age imputed was entered instead of mean age, the interaction term was still negative, but not statistically significant ($p = 0.677$).

Discussion

This meta-analysis has explored the effects of the use of rubrics on students' academic performance, self-regulation, and self-efficacy, along with moderating variables assumed to have an impact on these effects. The use of rubrics was shown to have a positive and moderate effect on students' academic performance, and a positive but smaller effect on students' self-regulation and

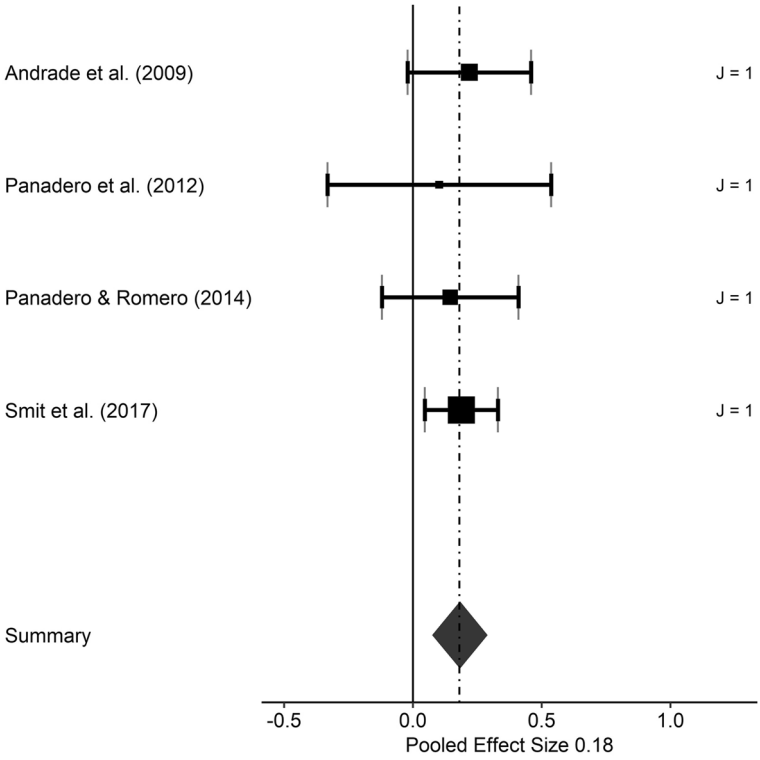

Fig. 9 Forest plot of studies reporting effects on self-esteem (standardized mean differences). Note. Black confidence intervals represent the global study-precision, whereas gray confidence intervals have been calculated using only sample size information. Therefore, the width of the gray confidence intervals is a direct function of the sample size, whereas the width of the black confidence intervals is a direct function of the global precision of the study, which depends on the sample size, the within study variability in that study, and the number of effect sizes reported. The dimensions of the squares correspond to the weights assigned to each study when computing the pooled effect size (larger squares indicate higher assigned weight). The J index denotes the count of effect sizes reported within the studies

self-efficacy. In most analyses, the potential moderator variables were not significantly related to the observed effect sizes.

The Effect of Rubrics and Moderators on Students' Academic Performance

In their review on the use of rubrics, Jonsson and Svingby (2007) claimed that it was not possible to draw any conclusions about student improvement related to the use of rubrics, as the findings were mixed. Similarly, Brookhart and Chen (2015) concluded that the "body of evidence that is accumulating is promising but not sufficient for establishing that using rubrics cause increased performance" (p. 363). The results from our meta-analyses, however, clearly lend support to the

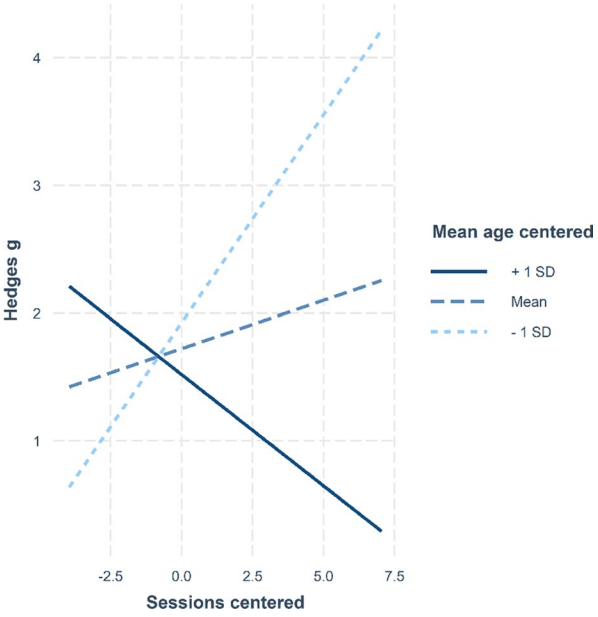

Fig. 10 Effect of the interaction between mean age and number of sessions (both variables centered)

conclusion that the use of rubrics aid students during task performance, so that they may perform at a higher standard as compared to control groups and/or on assignments performed before they were provided with rubrics.

The strength of the effect of using rubrics on students' academic performance was estimated by calculating Hedges' *g* using 54 standardized mean differences from 21 studies, which showed an overall effect of 0.57. However, as a slight asymmetry was detected in the funnel plot, a selection model was used to obtain a corrected overall effect size, providing different effect sizes for moderate and severe bias respectively. Since the Trim and Fill test did not detect the presence of publication bias, a small to moderate bias is more likely, suggesting an effect size of 0.45. Furthermore, according to calculations made from 30 standardized mean changes in 12 studies reporting on pre- and post-measures from both the intervention and control group, the effect was 0.38. The effect of using rubrics can therefore be assumed to be positive and moderate.

As suggested by previous research (Jonsson & Svingby, 2007; Panadero & Jonsson, 2013), our results show that we can keep the hypothesis that by making criteria and performance levels explicit, students can use them to regulate their learning and improve their performance.

Recent studies have looked at how different rubric designs might influence student performance. However, these studies, including our own, found that the specifics of rubric design do not significantly affect the outcomes (Jonsson & Panadero, 2017; Brookhart, 2018). This finding is somewhat surprising because previous researchers, like Jonsson and Panadero (2017), believed that carefully designing rubrics could enhance student learning. They suggested using multiple quality levels

and a detailed scoring method, rather than just a total score. Rubrics, by definition, include criteria and descriptions of performance levels (Brookhart, 2018). Most research on rubrics, therefore, already considers these elements in their design. The main differences in these studies are in the number of criteria and performance levels used, but these differences do not seem to have a major impact on how students perform. Brookhart (2018) also found no link between the type or quality of a rubric and student performance.

From the analyses of potential moderating variables, only mean age and educational level were found to have a statistically significant relationship to the observed effect sizes. Furthermore, this relationship was only seen when effect sizes were calculated from standardized mean changes (i.e., not standardized mean differences). Here, the effect of using rubrics was seen to decrease as the mean age of the sample increased (i.e., younger students benefited more from using rubrics). On the other hand, the effect for studies from secondary education was higher than the overall effect of studies from primary education and middle school, suggesting a reversed direction of effects as compared to students' age. Results from analyses of these moderating variables are therefore inconclusive, although the results from studies reporting on educational level may be more reliable, as the analyses on mean age were based on fewer effect sizes reported in two studies only, which make this conclusion very tentative. This would suggest that secondary students might benefit more from using rubrics as compared to younger students. However, there are no similar findings from earlier reviews for comparison. Previous reviews either concentrated on studies in higher education (Brookhart, 2018; Reddy & Andrade, 2010) or indicated a link between a student's age/educational level and the effectiveness of comprehensive interventions. It has been suggested that younger students might need longer and more extensive interventions for positive outcomes (Panadero & Jonsson, 2013). This theory aligns with our study's results, which demonstrate that intervention length has a lesser effect on older students. In contrast, studies in higher education frequently report positive outcomes from using rubrics, irrespective of intervention length (Brookhart, 2018; Panadero & Jonsson, 2013).

Student gender is a potential moderating variable that has been discussed in previous reviews, where the use of rubrics has been shown to have different effects on boys and girls in some studies, but not significant in others (Panadero & Jonsson, 2013). Brookhart and Chen (2015) suggest that there may be a gender effect in relation to writing, where girls seem to have an advantage, but whether there is a more general effect is unknown. The results from the current study suggest that there is no such general effect of gender on student performance from using rubrics in the studies included here.

Other potential moderating variables are self- and peer assessment. Jonsson and Svingby (2007) wrote that, although they found only few studies on this topic, these studies indicated that "rubrics might be valuable in supporting student self- and peer assessment" (p. 139). If rubrics facilitate self- and peer assessment, this may in turn support student performance (cf. Andrade, 2019; Yan et al., 2022). Panadero and Jonsson (2013) also suggest that there is evidence supporting the claim that rubrics may aid in improving student performance if combined with "self-assessment or other meta-cognitive activities" (p.

140). Our results show that the use of rubrics produced similar effect sizes if combined with either self or peer assessment. As self and peer assessment usually involve the use of metacognitive and regulatory strategies, both seem to be equally effective when combined with rubrics. It was not possible to empirically compare the effects of using self- and peer assessment against a control, however, as all studies included one or both strategies in combination with the use of rubrics.

Taken together, no claims can be made from this study beyond the conclusion that the use of rubrics have a positive and moderate effect on students' academic performance. It is important to note, however, that the exploration of some of the moderators is based on a narrow selection of studies, which might affect the stability of the results.

The Effect of Rubrics and Moderators on Students' Self-Regulated Learning

To investigate the effect of the use of rubrics on students' self-regulated learning, 17 effect sizes from five studies were synthesized. The overall effect was 0.23, which is considered a small effect. None of the moderator variables significantly explained the observed variability. Nevertheless, our study contributes by showing that there is indeed a positive effect of using rubrics on students' self-regulated learning, suggesting that rubrics can be used as a facilitator for students' self-regulation.

As noted by Panadero and Jonsson (2013), rubrics have been shown to facilitate several aspects of self-regulated learning. Improved self-regulation may therefore be a potential outcome of using rubrics. However, self-regulated learning is also a potential moderator, as students may use self-regulation strategies together with rubrics to improve their performance (see e.g., Jonsson, 2020). Complicating the matter further, self-regulation models comprise a wide range of cognitive, motivational, behavioral, and contextual variables, making Brookhart and Chen (2015) suggest that research on the relationships among these variables and rubrics still have a long way to go before we fully understand them. Unfortunately, the number of included studies did not allow for conducting specific analyses to contrast different self-regulatory strategies (e.g., motivation, emotions), thus one meta-analysis was performed for all effect sizes. Therefore, it can be recommended as a future direction of research to study more closely which parts of the self-regulation process rubrics can facilitate.

The Effect of the Use of Rubrics on Students' Self-Efficacy

Panadero and Jonsson (2013) discuss the influence of using rubrics on students' self-efficacy with reference to a study by Andrade et al. (2009), where the self-efficacy ratings of boys and girls are affected differently by long-term rubric use. Brookhart and Chen (2015) include a couple more studies in their discussion on rubrics and self-regulation of learning, but note that the evidence is mixed, where

some studies reported increased student self-efficacy from the use of rubrics, while others reported non-significant effects.

In our study, the overall effect of the use of rubrics on students' self-efficacy was small (0.18). However, this estimate is based on only four effect sizes from four different studies, which means that the result should be interpreted with caution. Still, it is interesting to note that this set of studies is particularly strong in terms of research design (i.e., either RCTs or well-planned quasi-experiments), they have substantial sample sizes (up to 762 participants in one of the studies), and some of the studies have control variables (e.g., self-regulated learning, different types of feedback as independent variables). Nevertheless, we still need to be careful with the interpretation of these results.

By providing transparency to the assessment situation, through criteria and performance levels, students could be expected to be in a better position to decide whether they are able to handle the assignment or not. If they think that they can, self-efficacy should increase, as should the use of productive self-regulation strategies. This is also the picture that emerges from studies investigating students' perceptions of using rubrics, such as the studies by Andrade and Du (2005), Jonsson (2014), and Reynolds-Keefer (2010) mentioned above.

However, it is also possible that the rubric could make the assignment appear more complex or difficult to complete (in particular with high standards), which would lower students' self-efficacy and potentially increasing the use of more negative self-regulation strategies, such as avoidance goals. This aspect is not well researched, although Panadero et al. (2013) report that the use of rubrics in fact *decreased* such negative self-regulation strategies. Similar to what was suggested in relation to self-regulation strategies above, the relationship between the use of rubrics and self-efficacy can be recommended as a future direction of research.

The Influence of Other Moderating Variables

In addition to the moderating variables discussed above, factors relating to the quality of research and publication were also explored, namely: (1) year of publication, (2) research design, and (3) the quality of the study. However, none of these were significantly related to the observed effect sizes. Since the range of research designs is very limited, encompassing only experimental and quasi-experimental studies, the lack of influence of this factor could perhaps be expected. Still, although the estimated quality among the studies provided greater variability (observed range was 5–13 points on a scale from 0–14), this factor did not seem to influence the outcomes either. Finally, it was assumed that later studies could have learnt from previous ones, resulting in larger effect sizes, but this assumption could not be confirmed.

Future Lines of Research

First, while the results from this meta-analysis are important for estimating the effects of the use of rubrics on students' academic performance, self-regulation,

and self-efficacy, the selected studies do not clarify the mechanisms for how rubrics influence students' academic performance, self-regulation, and self-efficacy. Consequently, research is needed that investigates more closely how students use rubrics by using external data sources (e.g., eye-tracking, direct observation), along with stronger measures of students' own perceptions and statements, such as thinking aloud protocols or validated self-report tools.

Secondly, more research is needed to understand the relationship between the use of rubrics and students' self-regulated learning. This relationship is complex, not least since self-regulated learning is multifaceted, encompassing cognitive, motivational, behavioral, and contextual aspects, which can all interact with the use of rubrics. As already noted by Brookhart and Chen (2015), research in this area has only just begun to explore these intricacies and still has a long way to go.

Thirdly, for future research to have a bigger impact in what we know about rubrics, the quality of the reports needs to improve dramatically. It is almost a constant in assessment research that the quality of reports is low, as mentioned in other reviews (e.g., Panadero et al., 2023). In the context of rubrics, there is often limited information available regarding two crucial aspects: rubric design and implementation. Access to details about rubric design reflects the quality of the instrument, while additional information on implementation sheds light on how teachers and students actually used the rubric. This information is essential for enhancing our understanding of the effects of rubrics. With the purpose of improving the quality of rubric studies reports we have created an instrument (see Appendix 4) following the original idea from Panadero and colleagues (2023). Our intention is that researchers could include this instrument as supplementary material to their articles. By having that information, future readers and colleagues conducting reviews or replications will be able to better understand the research design and the intervention. The instrument is freely available online via: https://osf.io/3hgcv/?view_only=31b7b778c81a40a5a7a00b55e416ad94.

Fourth, it will be important to conduct studies that investigate the long-term effects of rubric interventions. Such research could offer insights about, for instance, the sustainability of rubrics interventions, or whether there is transfer of students' knowledge about how to use rubrics from one subject to other subjects with different teachers.

Educational Implications

The major educational implication from this meta-analysis is that the use of rubrics has a positive influence on students' academic performance, and to a lesser extent on students' self-regulation strategies and self-efficacy. According to the analyses of potential moderating variables, it does not seem to matter how many criteria or quality levels the rubrics have, as long as they meet the basic definition of rubrics, and there are no clear indications of rubrics being more or less effective for younger or older students, or for students of different genders.

Conclusion

The use of rubrics has shown a positive and moderate effect on students' academic performance, and a positive but smaller effect on self-regulation strategies and self-efficacy. These findings might not appear as particularly striking, as there are already a number of systematic reviews drawing similar conclusions. However, the meta-analytical approach has the additional benefit of quantifying such effects. The observed effects make sense, since well-designed rubrics draw students' attention to the core aspects of how to execute the task with perfection. Furthermore, rubrics are used to evaluate whether students have indeed improved their performance, possibly reenforcing the effect even further. The finding that rubrics sometimes do not contribute to positive gains might therefore be seen as more startling.

To further our understanding of the processes underlying the effects, we explored the role of potential moderators. In most analyses, none of these potential moderator variables were significantly related to the observed effect sizes. An exception is that the length of the intervention was shown to have less impact on the observed effect in samples with older students, which means that older students, primarily in a higher education context, need less support to reap the benefits of using rubrics for formative purposes. It should be noted, however, that the lack of statistical significance in this analysis do not rule out a possible relationship between moderators and effects. In particular, there are two aspects of rubric research that limit the analysis of potential moderators. First, the number of studies exploring different moderators is relatively small, limiting the analysis. Second, there is quite some room for improvement when it comes to reporting important aspects of rubric design and implementation. To amend this situation, we are providing an instrument to support the report on rubric characteristics in future studies Appendix 4.

In conclusion, we assert that a key question to be answered by upcoming research is not whether rubrics work or not, but under what circumstances. In order to answer this question, future primary research need to clearly report on the specifics of the interventions, while future research reviews need to explore the characteristics of high-quality studies reporting on the successful implementation of rubrics, such as those identified in this meta-analysis.

Appendix 1

Method for the calculation of rubric presence in literature (Figure S1)

Search strategy

The search strategy consisted in employing the term Rubric* in WoS database. Specifically, we selected the "advanced search" option and then, we performed three independent searches with the same term but choosing different field tags: (1) Title, (2) AB (Abstract) and TS (Topic) (See Figure S1).

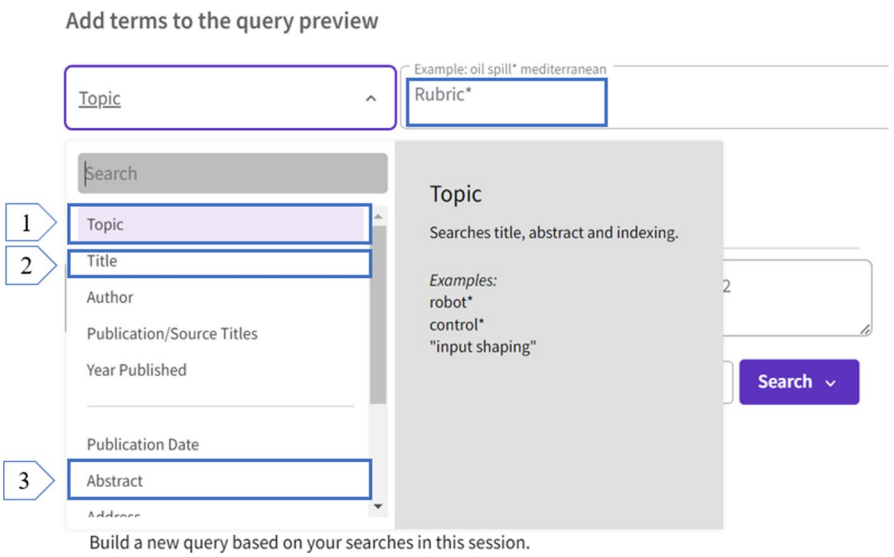

Fig. S1 Preview of the WoS search

Calculation process

Once we obtained the general results in each independent search (Title, Abstract and Topic), we filtered them by year. The specific steps were: in "Refine results" section, select "see all" and a new table will be open with the search records per year (Figure S2). We registered the year of publication and the total number of records of each year in an Excel. Finally, we created a Figure by using the total records data with Excel software.

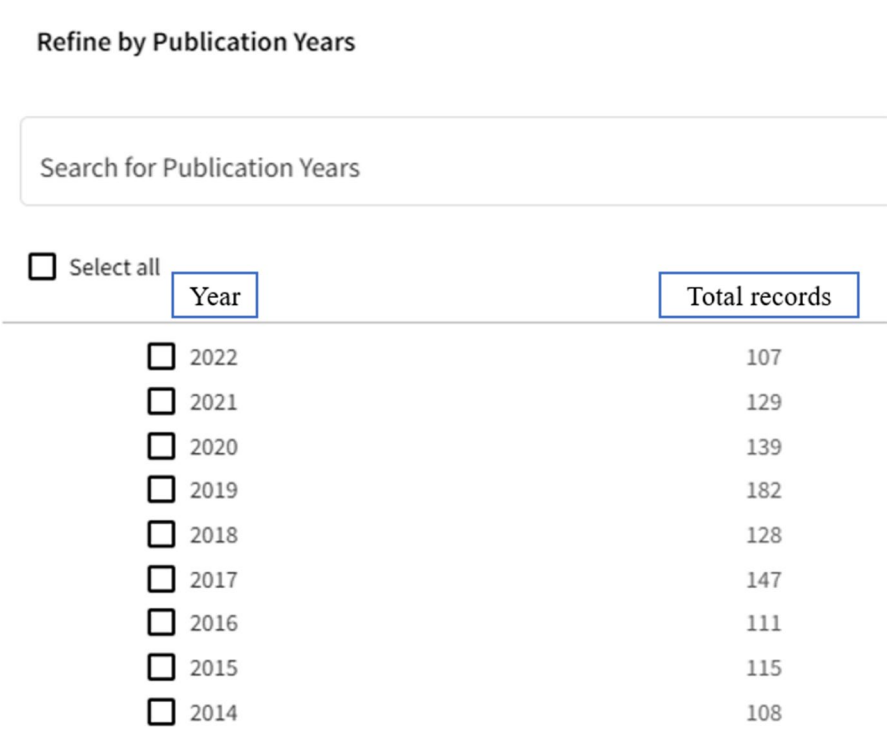

Fig. S2 Refine results section with publications per year

Appendix 2 Coding for quality of studies. Created by Panadero et al. (2023) with small changes to adapt it for our use

Coding categories		Weak (0 point)	Medium (1 point)	Robust (2 points)
Research design		Descriptive Design with Structured Observation Or Multiple Case Study	Quasi experimental Experimental design Intact Group ERIG	Experimental Experimental Design Random Group ERG
Quality of measurement tools		The study uses ad-hoc instruments without reporting reliability indexes	The study uses ad-hoc instruments reporting reliability indexes Performance is measured via classroom task designed by the teacher	The study uses validated instruments only Performance is measured via validated tools (e.g., nationwide test, institutional exams, etc.)
Quality of the intervention	Training	No training	Simple (e.g., explaining the rubric)	Advanced (e.g., feedback on the use of the rubric)
	Practice (Times the rubric was used)	1	2–5	6 or +
	Length of the intervention	Short (1 or two sessions)	Medium (3 to 6 sessions)	Long (months, semester, academic year)
Quality of DVs		e.g., Students' perceptions of…	e.g., SRL measured by self-report	e.g., "Objective" variables (academic performance, etc.)
Quality of the sample		Characteristics and selection process of the sample are not clearly described, or the sample is too small to extract conclusions	The sample is well-sized, but the participants were chosen using convenience-criterion only	The sample is well-sized, and the participants were chosen randomly or based in specific criteria other than convenience

Appendix 3

Publication Bias Analyses and Results of the Meta-regressions for Academic Performance Using As Effect Size the Standardized Mean Change

The selection method of Vevea and Woods (2005) estimates an overall effect size, corrected by publication bias. To do so, the selection method uses a weight function, in which "probabilities of being published" are assigned to each of the observe effect sizes based on their associated *p*-values, typically assuming that effect sizes associated with smaller *p*-values are more likely to be published (so the probability of being published is high) than effect sizes associated with higher *p*-values are less likely to be published (so the probability of being published is low). These probabilities are specified by the researchers. In our case, we specified two weight functions: one for moderate bias and another one for severe bias. In the weight model for moderate bias, we assigned a probability (of being published) of 1 to those effect sizes that were statistically significant ($p < 0.05$), a probability of 0.75 to those effect sizes whose associated *p*-values were between 0.05 and 0.50, and a probability of being published of 0.50 to effect sizes associated with *p*-values higher than 0.50. In the weight function for severe bias, we assigned a probability (of being published) of 1 to those effect sizes that were statistically significant ($p < 0.05$), a probability of 0.60 to those effect sizes whose associated *p*-values were between 0.05 and 0.50, and a probability of being published of 0.30 to effect sizes associated with *p*-values higher than 0.50.

Table S1 Results from the moderator analyses for **academic performance** for standardized mean changes

	m	*g* (SE)	*t* (*p* -value)	95% CI	$\delta^2_{between}$	δ^2_{within}
Publication year					0.171 (12)	0.100 (30)
Intercept	30	0.415 (0.161)	2.58 (.032)	[0.05, 0.78]		
Year[c]		0.035 (0.027)	1.27 (.312)	[-0.06, 0.13]		
Gender					0.160 (8)	0.101 (24)
Intercept	24	0.274 (0.179)	4.78 (.188)	[-0.19, 0.74]		
Perc. women[c]		0.010 (0.011)	1.79 (.498)	[-0.04, 0.06]		
Mean age					0.026 (2)	0.088 (14)
Intercept	14	0.514 (0.021)	21.1 (.021)	[0.29, 0.74]		
Mean age[c]		-0.060 (0.004)	-15.8 (.036)	[-0.10, -0.02]		
Mean age (imp)					0.125 (9)	0.086 (26)
Intercept	26	0.337 (0.146)	2.31 (.067)	[-0.03, 0.71]		
Mean age imp.[c]		0.017 (0.010)	1.69 (.198)	[-0.02, 0.05]		
Educational level					0.170 (12)	0.096 (30)
Primary	7	-0.007 (0.129)	-0.05 (.963)	[-0.42, 0.41]		
Middle	6	0.960 (0.169)	5.70 (.111)	[-1.18, 3.10]		
Secondary	1	0.138 (-)	-	-		
Higher	16	0.556 (0.273)	2.04 (.110)	[-0.20, 1.31]		

Table S1 (continued)

	m	*g* (SE)	*t* (*p* -value)	95% CI	$\delta^2_{between}$	δ^2_{within}
		Sig. difference between secondary ($d = 0.96$) and middle ($d = 0.14$, $p = .004$)				
		Sig. difference between secondary ($d = 0.96$) and primary ($d = 0.03$, $p = .004$)				
Type of education					0.234 (12)	0.089 (30)
Compulsory	14	0.209 (0.208)	1.01 (.361)	[-0.33, 0.75]		
Higher educati.	16	0.576 (0.277)	2.08 (.103)	[-0.18, 1.33]		
		There are not statistical differences between categories				
Number sessions					0.123 (8)	0.089 (25)
Intercept	25	0.582 (0.327)	1.78 (.171)	[-0.44, 1.61]		
Sessions[c]		-0.050 (0.034)	-1.49 (.206)	[-0.14, 0.04]		
Performance levels					0.248 (12)	0.095 (30)
4	18	0.372 (0.198)	1.77 (.116)	[-0.12, 0.86]		
5	12	0.416 (0.365)	1.98 (.245)	[-1.09, 1.92]		
6	0	-	-	-		
		There are not statistical differences between categories				
Assessment criteria					0.296 (10)	0.105 (27)
Intercept	27	0.410 (0.198)	2.07 (.074)	[-0.05, 0.87]		
Assessment[c]		0.004 (0.038)	0.10 (.930)	[-0.15, 0.16]		
Use of SA, PA or both					0.268 (12)	0.096 (30)
Self-assess.	24	0.435 (0.371)	0.71 (.607)	[-4.45, 4.98]		
Peer-assess.	5	0.264 (0.210)	2.07 (.077)	[-0.06, 0.93]		
Both	1	0.138 (-)	-	-		
		There are not statistical differences between categories				
Research design					0.221 (12)	0.100 (30)
Experimental	12	0.536 (0.297)	1.80 (.196)	[-0.60, 1.67]		
Quasi-experi.	18	0.309 (0.202)	1.52 (.173)	[-0.17, 0.79]		
		There are not statistical differences between categories				
Empirical quality					0.213 (12)	0.088 (30)
Intercept	30	0.361(0.166)	2.17 (.059)	[-0.02, 0.74]		
Quality[c]		-0.091 (0.073)	-1.25 (.260)	[-0.27, 0.09]		

Upper script *c* means that the variable has been centered prior the analyses. = within-study variance; = between-studies variance, *SA* self-assessment, *PA* peer-assessment

Appendix 4

Instrument to report the characteristics of rubric design and implementation

Purpose This instrument has been developed to provide comprehensive insights into the aspects of educational or research interventions utilizing rubrics. Furthermore, it serves as a tool to facilitate the design and implementation of more effective interventions. In order of relevance, the primary audiences include researchers

and teachers. Researchers can employ this instrument to enhance the clarity of their published papers, aiding fellow researchers in understanding the intervention's characteristics and study design. Teachers, on the other hand, can harness this tool to design more effective interventions and to report the specifics of their rubric-based instructional settings. Lastly, teacher educators and policymakers can benefit by gaining a broader understanding of the essential considerations when working with rubrics, enabling the development of stronger professional development courses for teachers and regulations that take these aspects into account.

How to Use it We envision three primary applications, though this list is not exhaustive.

1. **Designing Interventions**: Employ the instrument during the intervention design phase for studies, instructional settings, or policy development. Researchers, for example, can navigate the various aspects of the instrument to make informed decisions on how best to implement rubrics in their study. Teachers can use it to design their classroom interventions.
2. **Reporting Characteristics**: Use the instrument to document the characteristics of a rubric-based intervention, whether it's a scientific study, pedagogical documentation, or a guide for practitioners. The responsible individuals behind the intervention should complete the instrument and include it alongside other relevant documents. For instance, in a scientific publication, reference the instrument within the main body of the paper and attach it as an appendix for accessibility to other researchers.
3. **Systematic Reviews and Meta-Analyses**: The instrument can also be used for systematic reviews and meta-analyses. For example, the instrument could be included as an appendix in publications, so that the researchers conducting the review may compile those appendixes and, in this way, conduct more precise reviews and meta-analyses. Another example can be for researchers conducting a review to fill out the instrument for the studies they include, in order to have a clearer and more transparent coding of the rubric intervention characteristics.

Other Instruments of Interest: Dr. Ernesto Panadero and colleagues (2023) have previously developed an instrument for reporting the characteristics of peer assessment interventions. You can access it here: Link to Peer Assessment Instrument. Additionally, Dr. Panadero is working on another instrument to report the characteristics of self-assessment interventions, available upon request.

Instrument to report the characteristics of rubric design and implementation

Created by 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*.
Use this citation if you include the instrument in a publication.

Our study investigates:

- ☐ Rubrics and scoring accuracy
- ☐ Rubrics and academic performance
- ☐ Rubrics and students' perceptions
- ☐ Rubrics and_____

Describe the characteristics of your rubric intervention study in the table below.

	Category	Description	Our study
Design			
1	**Rubric presence**	Have you included the rubric in the publication as supplementary material?	☐ Yes ☐ No. Reason: Click here to add text
2	**Assessment criteria**	Number of assessment criteria included in the rubric	Click here to add text
3	**Performance levels**	How many performance levels are included in the rubric? Also list the headings	Click here to add text
4	**Creation**	Was the rubric created for this study? If not, please indicate the original source	☐ Yes ☐ No
5	**Scoring strategy**	If the rubric contains an explicit scoring strategy, provide a brief description.	Click here to add text
6	**Type**	How was the assessment communicated to the students, holistic (i.e., as an overall assessment for all criteria or analytical (i.e., separately for all criteria assessed)?	☐ Holistic ☐ Analytical
7	**Type 2**	Was the rubric general (i.e., a general skill such as writing), task-generic (i.e., applicable to several similar tasks) or task-specific (i.e., only applicable to one particular task)	☐ General ☐ Task-generic ☐ Task-specific
Implementation			
8	**Self-assessment**	Was the rubric used for self-assessment?	☐ Yes ☐ No
9	**Self-scoring**	Was the rubric used to calculate a self-score?	☐ Yes, but the self-score was not included in the final grade. ☐ Yes, and the self-score represented _% of the final grade. ☐ No
10	**Peer assessment**	Was the rubric used for peer assessment?	☐ Yes ☐ No
11	**Peer score**	Was the rubric used to score a peer?	☐ Yes, but the peer score was not included in the final grade. ☐ Yes, and the peer score represented __% of the final grade. ☐ No
12	**Feedback**	Did the students receive additional feedback about their performance or on how they used the rubric?	☐ Yes, on both ☐ Only on their performance ☐ Only on how they used the rubric ☐ No If yes, could you describe the additional feedback characteristics? Click here to add text
13	**Official weight**	Did the activity assessed with the rubric count towards the students' grade?	☐ Yes, for a ____% of the total ☐ No ☐ Click here to add text
14	**Frequency**	How many times was the rubric used? (Once, twice, etc.)	Click here to add text
15	**Training**	Did the participants receive training about the rubric? If yes, describe the training and the specific moment in which they received it.	Click here to add text
16	**Revision**	Did learners revise their work after using the rubric?	☐ No ☐ Yes

17	**Extent of involvement**	How were learners involved in the rubric design and implementation?	☐ Students just received and used the rubric ☐ Students were allowed to make small changes to the rubrics ☐ Students made substantial changes ☐ Students co-created the rubric ☐ Other: Click here to add text
18	**Use of other instruments**	Were any additional instruments employed to further strengthen the intervention effects, or to make comparisons with the rubric? If so, please, explain the characteristics of those instruments	Click here to add text
19	**Technology**	Was any type of technology used for the design and/or the implementation of the rubric? If so, please provide the details	Click here to add text
Outcomes			
19	**Study Outcomes**	These variables are directly measured as outcomes of the rubric activity. Select all the options that apply to your study from the right column.	☐ **Beliefs & perceptions:** including perceptions of learning capacity to use the rubric (e.g., fairness, usefulness), metacognition and self-regulation, attitudes and beliefs (e.g., self-efficacy), teachers' perceptions/conceptions. ☐ **Emotions and motivation:** emotions experienced by learners (e.g., achievement emotions, social emotions, etc.) & motivational beliefs (e.g., learning motivation). ☐ **Performance:** academic/domain specific performance, achievement, improved draft/work (i.e., revision). ☐ **Skills:** quality of contribution to the group, professional behaviour, problem solving skills, work habits, interpersonal skills, metacognitive & self-regulatory skills. ☐ **Reliability of rubric:** consistency of rubric scores among different raters (e.g., several teachers). ☐ **Validity of rubric:** aspects related to testing the validity, such as content validity, comparing students and teachers' assessment, etc. ☐ **Other:** Click here to add text
Moderators/mediators			
20	**Moderators/mediators**	Variables that are not usually manipulated but are taken into account when investigating rubrics. Select the variables that have been explored in your study from the right column.	☐ **Gender:** of assessor/assessee. ☐ **Ability & Skills:** includes prior knowledge, prior performance, achievement level, GPA, finished high school, previous level of education, year of enrolment, etc. ☐ **Skills:** reviewing ability, computer skills, etc. ☐ **Age/grade level:** of assessor/assessee. ☐ **Other:** Click here to add text
The design of this tool is based on an instrument to report peer assessment design characteristics from: Panadero, E., Alqassab, M., Fernández Ruiz, J., & Ocampo, J. C. (2023). A systematic review on peer assessment: Intrapersonal and interpersonal factors. *Assessment & Evaluation In Higher Education*, 1-23. https://doi.org/10.1080/02602938.2023.2164884. The two last categories (19 and 20) based on Alqassab, M., Strijbos, J., Panadero, E., Fernández Ruiz, J., Warren, M., & To, J. (2023). A systematic review of peer assessment design elements. *Educational Psychology Review*. https://doi.org/10.1007/s10648-023-09723-7 If your intervention included peer assessment we recommend you also fill out that instrument and include it as a supplementary material in your publication. It can be found here: https://osf.io/5k42z/?view_only=c77740eca9ef44978e1ac47abcaeef7c			

Acknowledgements We thank Philip Dawson for sharing the method he used on his review from 2017 to calculate the number of rubric publications. Additionally, we thank Heidi Andrade, Susan Brookhart, Philip Dawson, and Juan Fraile, for providing feedback on our instrument to report the characteristics of rubric design and implementation. Finally, we thank some of the authors of the included publications for answering our requests for clarification.

Funding Research funded by (1) Spanish National R + D call from the Ministerio de Ciencia, Innovación y Universidades (Generación del conocimiento 2020), Reference number: PID2019-108982 GB-I00. Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Data Availability The dataset and supplementary materials are available at: https://osf.io/d6mry/?view_only=0db0f9c5617a492c84500c0a83a185b9.

Declarations

Conflict of Interest The authors declare no competing interests.

References

Andrade, H. L. (2019). A critical review of research on student self-assessment. *Frontiers in Education, 4*, 87. https://doi.org/10.3389/feduc.2019.00087

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

*Andrade, H. L., Du, Y., & Wang, X. (2008). Putting rubrics to the test: The effect of a model, criteria generation, and rubric-referenced self-assessment on elementary school students' writing. *Educational Measurement: Issues and Practice, 27*(2), 3–13.

*Andrade, H. L., Wang, X., Du, Y., & Akawi, R. L. (2009). Rubric-referenced self-assessment and self-efficacy for writing. *The Journal of Educational Research, 102*(4), 287–302.

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

Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance*. Corwin Press.

*Bay, E., & Kotaman, H. (2011). Examination of the impact of rubric use on achievement in teacher education. *The New Educational Review, 24*(2), 283–292.

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology, 41*, 257–278. https://doi.org/10.1111/j.2044-8317.1988.tb00901.x

*Becker, A. (2016). Student-generated scoring rubrics: Examining their formative value for improving ESL students' writing performance. *Assessing Writing, 29*, 15–24.

Bloom, A. J., & Hautaluoma, J. E. (1987). Effects of message valence, communicator credibility, and source anonymity on reactions to peer feedback. *The Journal of Social Psychology, 127*(4), 329–338.

*Bohlin, S. L. (2000). *Effectiveness of instruction in rubric use in improving fourth-grade students' science open-response outcomes* (Publication No. 304607287) [Doctoral dissertation, University of Massachusetts Lowell]. Available from ProQuest Digital Dissertations database.

*Bradford, K. L., Newland, A. C., Rule, A. C., & Montgomery, S. E. (2016). Rubrics as a tool in writing instruction: Effects on the opinion essays of first and second graders. *Early Childhood Education Journal, 44*, 463–472.

*Brewer, D. (2002). *Teaching writing in science through the use of a writing rubric* (Publication No. 1760593502) [Doctoral dissertation, University of Michigan-Flint]. Available from M library. https://hdl.handle.net/2027.42/117681

Broad, B. (2003). *What we really value: Beyond rubrics in teaching and assessing writing*. University Press of Colorado.

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

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

Brookhart, S. M. (2018). Appropriate criteria: Key to effective rubrics. *Frontiers in Education, 3*(22). https://doi.org/10.3389/feduc.2018.00022

Cheung, M. W.L. (2014). Modeling dependent effect sizes with three-level meta-analyses: A structural equation Modelling approach. *Psychological Methods, 19* (2), 211–229. https://psycnet.apa.org/doi/10.1037/a0032968

Coburn, K. M., & Vevea, J. K. (2019). Weightr: Estimating Weight-Function Models for Publication Bias. R package version 2.0.2.https://CRAN.R-project.org/package=weightr

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

Dawson, P. (2017). Assessment rubrics: Towards clearer and more replicable design, research and practice. *Assessment & Evaluation in Higher Education, 42*(3), 347–360.

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). *Factors in judgments of writing ability* (ETS Research Bulletin RB-61-15). *Educational Testing Service*. https://doi.org/10.1002/j.2333-8504.1961.tb00286.x

Duval, S., & Tweedie, R. (2000). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*, 89–98.

Fernández-Castilla, B., Declercq, L., Jamshidi, L., Beretvas, N., Onghena, P., & Van den Noortgate, W. (2020). Visual representations of meta-analyses of multiple outcomes: Extensions to forest plots, funnel plots, and caterpillar plots. *Methodology, 16*(4), 299–315.

Fernández-Castilla, B., Aloe, A. M., Declercq, L., Jamshidi, L., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2021a). Estimating outcome-specific effects in meta-analyses of multiple outcomes: A simulation study. *Behavior Research Methods, 53*, 702–717.

Fernández-Castilla, B., Declercq, L., Jamshidi, L., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2021b). Detecting selection bias in meta-analyses with multiple outcomes: A simulation study. *The Journal of Experimental Education, 89*(1), 125–144.

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.

*Francis, J. E. (2018). Linking rubrics and academic performance: An engagement theory perspective. *Journal of University Teaching & Learning Practice, 15*(1), 3.

*Giacumo, L. A., & Savenye, W. (2020). Asynchronous discussion forum design to support cognition: Effects of rubrics and instructor prompts on learner's critical thinking, achievement, and satisfaction. *Educational Technology Research and Development, 68*, 37–66.

*Giacumo, L. A., Savenye, W., & Smith, N. (2013). Facilitation prompts and rubrics on higher-order thinking skill performance found in undergraduate asynchronous discussion boards. *British Journal of Educational Technology, 44*(5), 774–794.

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

*Greenberg, K. P. (2015). Rubric use in formative assessment: A detailed behavioral rubric helps students improve their scientific writing skills. *Teaching of Psychology, 42*(3), 211–217.

Hattie, J. (2012). *Visible learning for teachers*. Routledge.

*Hawthorne, K. A., Bol, L., & Pribesh, S. (2017). Can providing rubrics for writing tasks improve developing writers' calibration accuracy? *The Journal of Experimental Education, 85*(4), 689–708.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*, 107–128. https://doi.org/10.2307/1164588

Hillegas, M. B. (1912). A Scale for the Measurement of Quality in English Composition by Young People. *Teachers College Record, 13*(4), 1–1. https://doi.org/10.1177/016146811201300411

Jönsson, A. (2014). Rubrics as a way of providing transparency in assessment. *Assessment & Evaluation in Higher Education, 39*(7), 840–852.

Jönsson, A. (2020). Rubrics as a tool for self-regulated learning. In P. Grainger & K. Weir (Eds.), *Assessment Rubrics in Higher Education* (pp. 25–40). Cambridge Scholars Publishing.

Jönsson, A., & Panadero, E. (2017). The use and design of rubrics to support assessment for learning. In D. Carless, S. Bridges, C. Chan, & R. Glofcheski (Eds.), *Scaling up assessment for learning in higher education* (pp. 99–111). Springer.

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

Krebs, R., Rothstein, B., & Roelle, J. (2022). Rubrics enhance accuracy and reduce cognitive load in self-assessment. *Metacognition and Learning, 17*(2), 627–650.

*Lipnevich, A. A., Panadero, E., & Calistro, T. (2022). Unraveling the effects of rubrics and exemplars on student writing performance. *Journal of Experimental Psychology: Applied, 29*(1), 136–148. https://doi.org/10.1037/xap0000434

*Montanero, M., Lucero, M., & Fernández, M. J. (2014). Iterative co-evaluation with a rubric of narrative texts in Primary Education/Coevaluación iterativa con rúbrica de textos narrativos en la Educación Primaria. *Infancia y Aprendizaje, 37*(1), 184–220.

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

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

Panadero, E., & Jönsson, A. (2020). A critical review of the arguments against the use of rubrics. *Educational Research Review, 30*, 100329.

*Panadero, E., & Romero, M. (2014). To rubric or not to rubric? The effects of self-assessment on self-regulation, performance and self-efficacy. *Assessment in Education: Principles, Policy & Practice, 21*(2), 133–148.

*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.

*Panadero, E., Romero, M., & Strijbos, J. W. (2013). The impact of a rubric and friendship on peer assessment: Effects on construct validity, performance, and perceptions of fairness and comfort. *Studies in Educational Evaluation, 39*(4), 195–203.

Panadero, E., Jönsson, A., & Botella, J. (2017). Effects of self-assessment on self-regulated learning and self-efficacy: Four meta-analyses. *Educational Research Review, 22*, 74–98.

Panadero, E., Broadbent, J., Boud, D., & Lodge, J. (2019). Using formative assessment to influence self- and co-regulated learning: The role of evaluative judgement. *European Journal of Psychology of Education, 34*(3), 535–557. https://doi.org/10.1007/s10212-018-0407-8

Panadero, E., Alqassab, M., Fernández-Ruiz, J., & Ocampo, J. C. (2023). A systematic review on peer assessment: intrapersonal and interpersonal factors. *Assessment & Evaluation in Higher Education*, 1–23. https://doi.org/10.1080/02602938.2023.2164884

Popham, W. J. (1997). What's Wrong—and What's Right—with Rubrics. *Educational Leadership, 55*(2), 72–75.

Pustejovsky, J. (2021). clubSandwich: Cluster-Robust (Sandwich) Variance Estimators with Small-Sample Corrections. R package version 0.5.3. R package version 0.5.3. https://CRAN.R-project.org/package=clubSandwich

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

Reynolds-Keefer, L. (2010). Rubric-referenced assessment in teacher preparation: An opportunity to learn by using. *Practical Assessment, Research, and Evaluation, 15*(1), 8.

Sadler, D. R. (2014). The futility of attempting to codify academic achievement standards. *Higher Education, 67*, 273–288.

*Sáiz-Manzanares, M. C., Sánchez Báez, M. Á., Ortega-López, V., & Manso-Villalaín, J. M. (2015). Self-regulation and rubrics assessment in structural engineering subjects. *Education Research International, 2015*, 340521. https://doi.org/10.1155/2015/340521

*Smit, R., Bachmann, P., Blum, V., Birri, T., & Hess, K. (2017). Effects of a rubric for mathematical reasoning on teaching and learning in primary school. *Instructional Science, 45*, 603–622.

*Sundeen, T. H. (2014). Instructional rubrics: Effects of presentation options on writing quality. *Assessing Writing, 21*, 74–88.

Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019). A history of meta-regression: Technical, conceptual, and practical developments between 1974 and 2018. *Research Synthesis Methods, 10*, 161–179.

Torrance, H. (2007). Assessment as learning? How the use of explicit learning objectives, assessment criteria and feedback in post-secondary education and training can come to dominate learning. *Assessment in Education, 14*(3), 281–294.

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

Valentine, J. C. (2019). Incorporating judgements about study quality into research synthesis. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta—analysis* (pp. 129–140). Russell Sage Foundation.

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three- level meta-analysis of dependent effect sizes. *Behavior Research Methods, 45*, 576–594. https://doi.org/10.3758/s13428-012-0261-6

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2015). Meta-analysis of multiple outcomes: A multilevel approach. *Behavior Research Methods, 47*, 1274–1294. https://doi.org/10.3758/s13428-014-0527-2

Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods, 10*, 428–443. https://doi.org/10.1037/1082-989X.10.4.428

Viechtbauer, W. (2010). Conducting meta-analyses in R with the *metafor* package. *Journal of Statistical Software, 36*, 1–48.

Viechtbauer, W., & Cheung, M. W. L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods, 1*, 112–125.

Wen, M. L., & Tsai, C. C. (2006). University students' perceptions of and attitudes toward (online) peer assessment. *Higher Education, 51*(1), 27–44. https://doi.org/10.1007/s10734-004-6375-8

Wiggins, G. (1993). *Assessing student performance*. Jossey-Bass.

Wiggins, G. (1998). *Educative assessment*. Jossey-Bass.

*Wong, J. Y. H., Chan, M. M. K., Tsang, V. W. Y., Pang, M. T. H., Chan, C. K. Y., Chau, P. H., & Tiwari, A. (2021). Rubric-based debriefing to enhance nursing students' critical thinking via simulation. *BMJ Simulation & Technology Enhanced Learning, 7*(1), 11.

Yan, Z., Lao, H., Panadero, E., Fernández-Castilla, B., Yang, L., & Yang, M. (2022). Effects of self-assessment and peer-assessment interventions on academic performance: A pairwise and network meta-analysis. *Educational Research Review*, 100484. https://doi.org/10.1016/j.edurev.2022.100484