How accurate are our students? A meta-analytic systematic review on self-assessment scoring accuracy

Samuel P. León1*, Ernesto Panadero2 & 3, and Inmaculada García-Martínez4
1 Departamento de Pedagogía, Universidad de Jaén, Jaén, Spain.
2 Facultad de Educación y Deportes, Universidad de Deusto, Bilbao, Spain.
3 IKERBASQUE, Basque Foundation for Science, Bilbao, Spain.
4 Departamento de Educación, Universidad de Granada, Granada, Spain

Author note

Samuel P. León https://orcid.org/0000-0002-6980-2680
Ernesto Panadero https://orcid.org/0000-0003-0859-3616
Inmaculada García-Martínez https://orcid.org/0000-0003-2620-5779

Recommended citation: León, S. P., Panadero, E., & García-Martínez, I. (2023). How Accurate Are Our Students? A Meta-analytic Systematic Review on Self-assessment Scoring Accuracy. *Educational Psychology Review, 35*(4), 106. https://doi.org/10.1007/s10648-023-09819-0

Competing Interests: We have no known conflict of interest to disclose.

Research funded by: Second author participation funded by Spanish National R+D call from the Ministerio de Ciencia, Innovación y Universidades (Generación del conocimiento 2020), Reference number: PID2019-108982GB-I00.

Correspondence concerning this article should be addressed to Samuel Parra León, Paraje de las Lagunillas s/n, 23071 Jaén (España), e-mail: sparra@ujaen.es

How accurate are our students? A meta-analytic systematic review on self-assessment scoring accuracy

Abstract

Developing the ability to self-assess is a crucial skill for students, as it impacts their academic performance and learning strategies, among other areas. Most existing research in this field has concentrated on the exploration of the students' capacity to accurately assign a score to their work that closely mirrors an expert's evaluation, typically a teacher's. Though this process is commonly referred to as self-assessment, a more precise term would be self-assessment scoring accuracy. Our aim is to review what is the average accuracy and what moderators might influence this accuracy. Following PRISMA recommendations we reviewed 160 articles, including data from 29,352 participants. We analysed 9 factors as possible moderators: (1) Assessment criteria; (2) Use of Rubric; (3) Self-assessment experience; (4) Feedback; (5) Content knowledge; (6) Incentive; (7) Formative assessment; (8) Field of knowledge, and (9) Educational level. The results showed an overall effect of students' overestimation (g = 0.206) with an average relationship of z = 0.472 between students' estimation and the expert's measure. The overestimation diminishes when students receive feedback, possess greater self-assessment experience and content knowledge, when the assessment does not have formative purposes, and in younger students (primary and secondary education). Importantly, the studies analyzed exhibited significant heterogeneity and lacked crucial methodological information. Therefore, although our study shows that there is room for further research in the area, these studies should consider the conclusions of our study.

Keywords: self-assessment, self-grading, accuracy, bias, meta-analysis.

How accurate are our students? A meta-analytic systematic review on self-assessment scoring accuracy

Self-assessment is among the most investigated instructional / learning strategies in educational psychology as shown by, both, the historical recollection of studies and the number of reviews on the topic (e.g., Falchikov & Boud, 1989; Sitzmann et al, 2010). The reasons for its relevance are multiple. Firstly, self-assessment is the capacity to assess oneself which is, obviously, a key skill for students to master (Boud, 1995; Panadero et al., 2019; Tan, 2012). Additionally self-assessment has a positive effect on students' academic performance (Brown & Harris, 2013; Yan et al., 2023), is an essential component in formative assessment (Black & Wiliam, 1998), and plays a crucial role in the students' ability to self-regulate their learning (Cascallar et al., 2006; Panadero et al., 2017). Thus, its prevalence in the educational psychology literature.

The area that has traditionally garnered more focus within the field of self-assessment is the measurement of accuracy, typically evaluated through score accuracy (i.e., closeness of the learner's own score about her work compared to a yardstick usually the teacher's score) (Dochy et al., 1999; Ward et al., 2002). Even current reviews show that it is still the dominant line of work (Andrade, 2019). Importantly, the inception of formative assessment shifted a significant part of the attention away from grades towards the use of assessment practices to impact students' learning (Wiliam, 2011), turning self-grading accuracy into a disputed arena (Andrade, 2019). Nevertheless, self-assessment accuracy is still the most studied topic as shown by its massive presence in published work (Andrade, 2019) and probably based on the fact that this type of self-assessment might have pedagogical value to justify its exploration (Panadero et al., 2016; Sitzmann et al, 2010).

While there are two previous meta-analyses on self-assessment accuracy (Brown & Harris, 2013; Falchikov & Boud, 1989), we believe conducting a new meta-analysis is needed at this point for four reasons. First, many years have gone by since the last one for an area that is very highly published (see for example, Andrade, 2019). Second, the previous meta-analyses focused on specific educational levels, thus there is need for a meta-analysis comparing across levels. Third, since 2013 there have been important methodological and analytical advances in meta-analysis to better examine moderators and their role. And fourth, a more systematic analysis including new moderators would be beneficial. These aspects will be further elaborated upon in the forthcoming theoretical framework section.

All above considered, this study aims to examine the available empirical evidence regarding the accuracy of self-assessment and its effects, as well as to identify potential factors, commonly referred to as moderators, that may influence these effects.

Self-assessment and its accuracy

"Student self-assessment most generally involves a wide variety of mechanisms and techniques through which students describe (i.e., assess) and possibly assign merit or worth to (i.e., evaluate) the qualities of their own learning processes and products" (Panadero et al., 2016 p. 804). As shown by the same state-of-the-art review, self-assessment is quite heterogeneous taking many different forms, from simplistic self-grading approaches in which students award their work a grade with no further training or feedback to more complex ones such as "integrated tutor and peer feedback before self-assessment" (Taras, 2010). We adopted this definition as to include the maximum number of available studies on self-assessment accuracy.

Despite the various self-assessment approaches and the introduction of formative assessment, which has diminished the emphasis on grading, the accuracy of

self-assessment scoring, henceforth self-assessment accuracy, continues to be the most extensively researched aspect in the field of self-assessment (Andrade, 2019). As mentioned earlier, self-assessment accuracy can be defined as the evaluation of the proximity of the students' self-assessment to that of a comparator. Crucially, this is typically assessed by having students assign a grade to their work, which is then compared to a standard or benchmark (Ward et al., 2012). Scoring accuracy remains as the almost exclusive construct explored though there has been propositions for "content accuracy" that is the degree of closeness between the student's identification of rights and wrong and ways of improving the work when compared with the teacher's (Panadero et al., 2016). Additionally, it is worth noting that different terminologies have been proposed to refer to "scoring accuracy" such as "construct validity" (Panadero et al., 2013) or "consistency" (Andrade, 2018), but the underlying rationale is the same. In this discussion, we will adhere to the term "accuracy" due to its widespread usage. Finally, we will use the more general term of self-assessment, though most of the studies analysed here have a summative purpose and could be labelled as self-grading.

One crucial question is, does self-assessment accuracy have positive effects? Empirical research shows that it has. Sanchez and colleagues (2017) found that asking students to self-grade their work increased their performance ($g = .34$) on subsequent tests when compared with students that did not self-assess. Then, the next question being, is scoring self-assessment accuracy important? Panadero et al. (2016) argue that accuracy is key as students need to know what went wrong to correct it and what was correct as to repeat it. For that: "In SSA [Student Self-Assessment], such inferences depend, in part, upon SSA decisions being reasonably realistic. Realism, sometimes referred to as veridicality (Butler 2011), identifies the degree to which student descriptions about their work are perceptibly true or accurate" (p. 811). While these

authors argue that "content accuracy" might be more important than "scoring accuracy" for learning itself, they also believe that -due to the importance that scores have for students- scoring accuracy is still an opportunity for veridicality of the self-assessment performed and, ergo, influential for learning. Also, the research in other areas of psychology, points out about the importance of a calibrated self-assessment (e.g., Dunning et al., 2004; Gutierrez de Blume, 2022; Stone, 2000).

Finally, in a more nuanced way, Brown and Harris (2013) stated in their meta-analysis: "…self-assessment is not robustly accurate but also it certainly is not randomly related to external measures of performance. Correlations falling in the range of 0.30 to 0.50 explain some 10% to 25% of variance between the self-assessment and some external measure of performance" (p. 386). Therefore, from the academic performance point of view, it makes sense to explore the students' scoring self-assessment and its accuracy.

Are students accurate self-assessors?

To our knowledge, two meta-analyses have been published that explore how accurate are students in formal education when asked to self-assess. Firstly, Falchikov and Boud (1989) published the first meta-analysis about educational self-assessment. Their main aim was to calculate student's self-assessment accuracy when compared with teachers' scores. Their search produced 57 studies including 96 comparisons, with effect sizes ranging from -0.62 to 1.42 (negative values indicating under-estimation of the students, and positive values overestimation). The mean effect size was 0.47. Interestingly, they also found that there were wide variations in the percentage of students which self-assessment coincide with their teachers, from 33% to 99.4%. The average was 64.1%.

In the second meta-analysis, Brown and Harris (2013) explored the relationship of self-assessment to academic achievement, effects of self-assessment on self-regulation processes, student perceptions of self-assessment, and accuracy of self-assessment. For the latter, they also explored the role of student age, schooling experience, ability, task features and method of self-assessment. They reported positive correlations within a wide range ($r \approx 0.20$ to 0.80), but only a limited number of studies documented correlations exceeding 0.60. These results were found in three types of comparisons: (1) self-ratings and teacher ratings, (2) self-estimates of performance and actual test scores, and (3) student and teacher rubric-based judgments. Regarding the effects of age and ability these them to be confounded so it was not so straight-forward to disentangle which one produced the effects. Nevertheless, younger students tended to present overscoring while older ones were closer to the teachers' scores. Regarding ability, abler students were more accurate. Regarding task features, activities that were familiar and predictable for the students seem to produce more accurate self-assessment. Regarding method of self-assessment, the more specific and concrete the standards and criteria, the higher the accuracy. Finally, they also reported on other variables that could influence accuracy but with less certainty such as gender, ethnic culture, personality, or training. It would be important to explore what is the current knowledge about all the moderators explored in Brown and Harris (2013).

It can be then concluded that the previous empirical evidence has found a clear relationship between scoring self-assessment and teachers' grade and that students tended to overestimate their performance, especially when some moderators played a role. However, there are four reasons why further research is needed. First, the two previous meta-analyses were conducted years back and there is a need to investigate if the new evidence is in the same direction, especially knowing that accuracy is by far the

most studied self-assessment area even these days (Andrade, 2019). Second, Falchikov and Boud (1989) focused in higher education, while Brown and Harris (2013) focused on the K-12 sector. Therefore, to our knowledge there is not a meta-analysis exploring those educational levels together. This is important as in both meta-analyses the role of age and schooling level seems to play a role, and the direct comparison among those levels has not been previously done. Obtaining such knowledge would help understand what is the effect of educational experience and developmental stage on self-assessment accuracy, if any. Third, there has been important methodological and analytical advances in meta-analysis that could help us better understand what are the effects and the role of the moderators. Namely and shortly, there will be six main methodological advantages from the previous meta-analyses: (1) we will follow PRISMA protocol, (2) we will be able to calculate a more sensible effect size, (3) we will use a multi-level meta-analysis approach, (4) analysis of outliers and/or influential values, (5) analysis of publication bias, and (6) a detailed analysis of moderators effects. And fourth, regarding the effects of moderators, despite having been explored especially in Brown and Harris (2013), these need to be approached with a greater level of systematicity considering the decade of research since 2013. By doing so, we would be able to profile what are the main characteristics that could enhance the success of a self-assessment intervention.

Importantly, scoring self-assessment represents a heterogeneous strategy because it can follow different implementations strategies (e.g., the use of not of rubrics, delivering or not feedback about the accuracy of the self-grade) and, or because of this, the proximity of the self-grading to the teachers' score varies. This is an important aspect to review as mentioned earlier. Do some implementation variables –i.e., moderators- modify the accuracy of self-assessment?

The role of moderators in scoring self-assessment

Next, we present shortly what type of moderators have been found to influence self-assessment accuracy in previous research as to include them in our meta-analysis and explore their influence (Falchikov & Goldfinch 2000; Panadero et al., 2016; Ward et al., 2002).

First, when students self-assess they do it using *criteria*, which is usually personally established or led by the teacher (Butler & Winne, 1995; Panadero & Alonso-Tapia, 2013). If it is the last case, then it is to be expected that if they receive the criteria explicitly, in whatever form, this helps them to be more precise in their self-assessment (Klenowski, 1995). This was one of the main factors proposed by Falchikov & Boud (1989). The authors found that studies in which the evaluation criteria were presented and understood by the participants, prior to their self-assessment, showed more accurate self-assessments. Not knowing the evaluation criteria will cause students to evaluate more subjectively.

Second, *rubrics* have been used for some decades as a way to increase the accuracy and learning effects of self-assessment among other purposes (Brookhart, 2018). Actually, Andrade coined the strategy of rubric-referenced self-assessment (Andrade & Du, 2005) that shows the interrelation between rubrics and self-assessment. It is to be expected that the use of rubrics will enhance students' accuracy (Jonsson & Svingby, 2007). Assuming that providing students with evaluation criteria to reduce the subjectivity of their evaluations, the use of rubrics has been shown to be one of the best ways to facilitate those criteria for students (e.g., Krebs et al., 2022; Panadero & Romero, 2014).

Third, it could be expected that when students have *previous self-assessment experience* this would increase their accuracy. However, the two previous meta-analyses focus on the expertise on content knowledge or educational level (Brown &

Harris, 2013; Falchikov & Boud, 1989), not on the previous experience on self-assessment itself. Therefore, we will explore whether the studies explored here analysed the effect of previous experience in self-assessment.

Fourth, it is important to explore what is the role that *feedback* could play on self-assessment accuracy. It has been argued that external feedback is needed as to correct for bias in internal feedback / self-feedback (Butler & Winne, 1995; Panadero et al., 2019). It has also been argued that teachers can strengthen the effects of self-assessment through feedback (Boud, 1995), and that such feedback is even more powerful if covers not only the task but also the self-assessment itself (Panadero et al., 2016). However, when it comes to the specific line of research on how feedback on the scoring accuracy influence that same accuracy results are mixed with some studies showing a positive influence (Baars et al., 2014; Dunlosky et al., 2011; Lipko et al., 2009), and others the contrary (Raaijmakers et al., 2019). Thus, it will be important to clarify what does the bulk of empirical evidence indicate. We can give as an example of the two previous factors the study conducted by León & García-Martínez (2021). This study measured the accuracy of students' self-assessment over 9 sessions. The students were divided into two groups, one of which received feedback after completing their assessment and the other group did not. The results showed that all students improved their accuracy over the assessment sessions, but those who also received feedback improved their accuracy faster than those who did not.

Fifth, it has been argued and proven that higher level of *content knowledge* increases accuracy, a result found in the two previous meta-analyses (Brown & Harris, 2013; Falchikov & Boud, 1995). Actually, those authors identified that it was the content knowledge in a specific task or subject what made the difference not the seniority of the participants. A more recent example of this effect was found by León et

al., (2021) who found that accuracy varied significantly as a function of the level of proficiency, thus revealing biases associated with this level of proficiency. Thus, due to the importance of this variable, we included it as a moderator.

Sixth, previous studies have suggested that students' *motivation* to be accurate in their self-assessments might influence their accuracy (Panadero et al., 2016; Tejeiro, et al., 2012). To increase such motivation to be accurate, students have sometimes received incentives (e.g., some kind of bonus on the final mark) for realistic and accurate self-assessment (Abdalla, et al., 2021). Additionally, previous research has shown that when students have mastery goal orientation they increase their self-assessment practices (Yan, 2019). Nevertheless, the evidence is still limited, and we aim to analyze it here to further clarify this relationship.

Seventh, it has been argued that self-assessment when designed with a *formative purpose* could increase students' learning (see Harris & Brown, 2018 for an excellent overview). For this reason, there has been ample debate in the field on whether self-assessment should be only used for formative purposes as the summative purpose could distract the students from such learning (Andrade, 2018, 2019). Nevertheless, there is empirical evidence that students can benefit when they self-assess with summative purposes (e.g., Sánchez et al., 2017). Thus, we will analyse whether interventions with a formative purpose influence scoring accuracy (i.e., summative purpose) even if there is an obvious tension between the purpose and the outcome of the self-assessment.

Eight, it is also important to explore if the *field of knowledge/subject* influences the accuracy in self-assessment. We could answer questions such as whether performing self-assessment in mathematics leads to higher accuracy than doing it in English. Unfortunately, there is barely any research comparing in the same study different

subjects (Panadero et al., 2020), so performing a meta-analytical exploration is particularly important for this moderator.

Ninth, it is important to compare the *educational level* effects as it could be that older students might turn more accurate over time. However, as mentioned earlier, seniority is not always a significant moderator and there is an unclear distinction among age and level of schooling (Brown & Harris, 2013; Falchikov & Boud, 1989). As the two previous meta-analyses focused exclusively in one educational level, it is important to explore this moderator. Unfortunately, there are barely any previous studies exploring different educational levels due to the complexity of such approach (e.g., having secondary and higher education students doing the same task due to the differences in skills). We only know of a few studies that have chosen this approach (Panadero et al., 2020, 2022; Yan, 2018) but they did not explore scoring accuracy. Therefore, we will try to clarify through meta-analytic approach if there is a clear direction.

Finally, tenth, accuracy needs to be operationalized through a *comparator* that acts as the benchmark to evaluate the distance of the students' self-assessment to such comparator. This benchmark has been usually the teachers' grade (Dochy et al., 1999), though it has been debated it this is the ideal one and ways to strengthen the comparator (Ward et al., 2002). It is important to explore if teachers' grade is still the usual yardstick and to compare its effects against other comparators.

Aim and Research questions

Considering the characteristics of the two previous meta-analyses, the bulk of incoming research and the need for further clarify its impact, our aim is to investigate the empirical evidence around the students' self-assessment scoring accuracy and its moderators. We will explore the following research questions (RQs):

RQ1. What are the main characteristics of self-assessment accuracy studies?

RQ2. What is the average accuracy of students' self-assessment?

RQ3. Do the moderators influence students' self-assessment accuracy?

Materials and Methods

Search procedures

This systematic review follows the recommendations of PRISMA (Moher et al., 2009) and APA's (Appelbaum et al., 2018) reporting standards for systematic reviews and meta-analyses (for further details of the PRISMA process see Supplementary Table S1). On March 11th 2021, the first author performed an electronic search on the *Web of Science, ERIC,* and *PsycINFO* using the search terms "self-assessment" and "self-evaluation". These terms were chosen as they represent the wider ones as to try to identify the maximum number of articles that explore our intended area. On March 28th 2023, the search in Google Scholar was repeated to extend the search to the gray literature. This search was performed with the free software *Publish or Perish* (Harzing, 2007). The search was limited to a) peer reviewed English-language articles, b) published after 1989 -date on which Falchikov and Boud published their famous systematic review and meta-analysis on self-assessment (SA)-, c) with categories restricted to "education/educational research" and "psychology". Unpublished dissertations, reviews and meta-analyses were excluded at this stage. After eliminating 153 duplicates, the number of articles resulting from the initial search was 17,581 studies.

In the first filtering phase, the first author reviewed the title and abstracts of the 17,581 articles using the following inclusion criteria: 1) empirical quantitative studies where scoring self-assessment is implemented and could be compared with an expert's assessment (e.g., teacher, instructor, tutors,), 2) participants were students at any level of formal education, 3) the self-assessment was calculated by evaluating students'

academic performance on a task, 4) students and expert must have used the same measurement scale (e.g. 100 points, 10 points), 5) articles were written in English; and 6) they were peer-reviewed. This first screening resulted in a total of 356 full-text articles.

In the next step, the first and third authors independently read the full text of the articles in order to verify that they met the established inclusion criteria. As a result of this screening, a total of 131 articles were finally selected for inclusion in the study. Finally, the first author analysed the papers cited in the two previous reviews focusing on the accuracy of self-assessment in educational contexts (Harris & Bound, 2013; Andrade, 2019). Of the 243 papers filtered by title and abstract, 32 were analyzed by complete reading of the paper, and only 1 paper was finally selected for inclusion.

Figure 1 represents the PRISMA Flowchart of the literature search process conducted in this study. Across the 160 full-text articles read for inclusion, the inter-rater agreement was 97.88%. Cohen's Kappa was 0.95 indicating a near-perfect agreement between judges. Disagreements were resolved by discussion and consensus between the two researchers until a 100% agreement was reached. Table 1 shows a summary of the most relevant information from the articles included.

[Figure 1 near here]

Data extraction and coding

The first and third authors independently coded each of the 160 papers selected, including descriptive information on the papers, such as the aim of the study, size and age of the sample, the educational level and subject area, a summary of the design used, the dependent variable and test with which it was measured, and the type of analysis performed, a summary of the results (see Table 1).

In addition to the descriptive information in Table 1, relevant information related to the practice of self-assessment was collected (see Table 2). We will now

describe the criteria we adopted when coding the information for the factors associated with self-assessment practices. If the factor is dichotomous (Yes or No), we will indicate when we consider the presence of the factor to be met. Otherwise, it will be coded as No. The factors included in Table 2 were:

(1) Assessment criteria: provision of evaluation criteria for the self-assessment. We will consider this moderator to be met (Yes) when the study reports that the students have been informed or provided with the same evaluation criteria that will be used by the expert to evaluate performance on the test.

(2) Use of a rubric: whether this tool was provided or not. We will consider this moderator to be met (Yes) when before the students take the self-assessment test they are provided with the same rubric that will be used by the expert to evaluate the test.

(3) Self-assessment experience: previous experience. We will consider the moderator to be met (Yes) when students have had previous experience performing self-assessments. This experience may have been acquired either in self-assessment training assignments, or having done self-assessments in previous sessions.

(4) Feedback: whether it was delivered on the task that was performed. We will consider the moderator (Yes) to be met, when the study reports that the students receive feedback on their performance in the assessment. Any type of feedback will be considered as long as it comes from the expert and aims to inform the student of his performance in the test.

(5) Content knowledge: level of competence in the task being self-assessed. We will consider that this moderator (Yes) will be met when it is indicated that the students are knowledgeable about the subject to be evaluated. Thus, we

will consider that they are knowledgeable about the subject either when they are in higher stages of studies (e.g., post-graduate students or final year language students) or when in the study, students are grouped by their level of proficiency in the subject.

(6) Incentive: whether learners received incentives to be accurate in their self-assessment. We will consider that the moderator (Yes) is fulfilled, when the study reports that those students who manage to be accurate in their self-evaluations will have some type of compensation (e.g., on the grade, economic).

(7) Formative assessment: whether the potential formative capacity of the self-assessment was analysed. We will consider that the moderator (Yes) is fulfilled when the study analyses the possible formative effect of having carried out a self-assessment. Thus, any study that analyses whether the fact that students carry out a self-assessment has a positive impact on their learning of the analysed subject. Otherwise the moderator will be coded as No. It will not be considered when analysing whether self-assessment improves the accuracy of future self-assessments.

(8) Expert evaluator: type of expert who conducted the assessment that was used as yardstick to evaluate the accuracy of the self-assessment. We will record who is the agent that performs the functions of expert in the evaluation.

(9) Self-assessment purpose: the instructional aim for which the self-assessment was used in the study (e.g., accuracy, self-regulated learning).

Computation of effect sizes and statistical analyses

In addition to the qualitative information just presented, we also extracted quantitative information concerning students' self-assessment accuracy. Importantly, to calculate accuracy the selected studies had used two analytical techniques. First, the

contrast of means, where the scores given by the students were compared with the scores given by the experts. To compute the effect size in these studies, standardised mean difference test scores (g_p) were calculated. We used Equations 4.18 and 4.19 from Borenstein et al. (2009) including also the correction factor *J* computed with Equations 4.22 and 4.23 to calculate the effect size of this first type of studies. The variance of g_p was computed using Equations 4.20 and 4.24 from the same source. And the second analytical technique that was used in some of the selected studies was correlational analysis between students' and experts' scores. For this group of studies, instead of analysing the raw correlation coefficients (with values between -1 and +1 and non-normal distribution), we transformed the correlations to Fisher's *z* scores (Silver & Dunlap, 1987).

Consequently, our quantitative analysis is composed of two meta-analyses, one for studies in which we computed g_p and another Fisher's *z*. The g_p meta-analysis calculates the differences between student and expert estimates, thus effects with values close to 0 indicate good accuracy in self-assessment, positive effects indicate overestimation by students and negative effects underestimation. In the meta-analysis based on correlations, positive correlations close to 1 indicates a good fit between the students' estimates and the experts' assessments, and small correlations or correlations close to 0 indicate low accuracy in the students' self-assessment. Outliers were detected and removed in each of the meta-analyses.

Since some of the included studies offered more than one dependent variable, to analyse the statistical dependence of the effect sizes within the same study, we adjusted the analysis with a multi-level random-effects model using the rma.mv function of the *metafor* package (version 3.4.0) for R (Viechtbauer, 2010) using random-effects models. To identify studies with outlying outcomes, after adjusting the multilevel

model, Studentized residuals (> 2) and Cook's distance (> 4/n) were estimated (Viechtbauer & Cheung, 2010), those effect sizes identified as outliers were excluded from meta-analyses.

After estimating the average effects of the two proposed meta-analyses, we assessed the influence of the factors extracted from Table 1 (moderators 8 and 9) and Table 2 (moderators 1 to 7) as possible moderating variables of the effect sizes. Due to the lack of information on the factors type of expert assessor and purpose of the self-assessment, these two could not be included in the moderation analysis. The final variables for the moderation analysis were: (1) Assessment criteria; (2) Use of Rubric; (3) Self-assessment experience; (4) Feedback; (5) Content knowledge; (6) Incentive; (7) Formative assessment; (8) Field of knowledge, (9) Educational level.

To analyze the possible publication bias of the studies included in our review, we carried out two approaches. First, we analyzed the asymmetry of the effect sizes represented in the Funnel Plot through Egger's test. In the second approach we analyzed the distribution of the p-values in the studies that reported significant effects through p-curve analysis. *P*-curve analysis is an analytical method developed to detect the possible existence or absence of an effect by analysing the distribution of significant *p*-values reported in studies (Simonsohn et al., 2014). When faced with a true effect, the distribution of significant *p*-values should be significantly more present below *p* .25 than those between .025 and .05. This analysis was carried out using the web-based application developed by Simonsohn and colleagues (available at http://www.p-curve.com/app4/)

Results

RQ1. What are the main characteristics of self-assessment accuracy studies?

Table 1 summarises the information of the 160 papers included in the review. As

can be seen in Figure 2, the number of papers that explores self-assessment accuracy has increased dramatically over the years. The total sample across all included papers was 29,352 participants. The included studies covered educational stages from primary education to post-graduate students, but some studies did not report the age of the participants and the actual range that was explicitly presented was between 12 and 46 years old. The fields of knowledge with the highest presence were Medicine (33.12% of all studies), Languages (10.62%) and Psychology and Sciences (5.62%). The predominant educational level of studies was higher education (79.37%), followed by secondary education (7.5%). The most commonly used analytical techniques in the included studies were correlation and contrast of means through analysis of variance. Regarding the procedure used, a great heterogeneity was found. In general, the most employed procedure was that after a learning phase, the students had to carry out an evocation test of what they had learned, which was evaluated by the expert, and additionally the students had to carry out a self-assessment test. As in the case of the design, or maybe because of it, there was a great deal of heterogeneity in the results found. For more detailed information, see Table 1.

[Insert Table 1 near here]

[Figure 2 near here]

Table 2 shows information related to the type of self-assessment that was performed in each of the included studies. Of all the studies included, 63.13% of the students knew about the assessment criteria, 33.13% were not aware of them and in 3.75% of the studies there were students who were aware of them and others who were not. In 56.25% of the included papers a rubric was used in the evaluation, while in 43.75% it was not. Regarding experience in conducting self-assessments, in 51.25% of

the studies the students had no experience, in 15.00% they had experience, and in 33.75% this condition was manipulated throughout the study. 55.63% of the assignments did not offer feedback to students after the assessments, 38.13% did, and 6.25% offered feedback to some students and not to others. In relation to content knowledge, in 16.67% of the studies the students had knowledge of the subject to be evaluated, in 52.27% they did not, and in 15.91% this factor was manipulated between or within groups. Few studies looked at self-assessment accuracy when the learner was motivated to be accurate (12.50%). Of the total number of studies, only 20.63% analysed the possible formative effect of self-assessment, while 78.75% did not assess it. Only one study (0.63%) analysed this factor between groups. Of the information recorded on the expert assessed, 75.63% of the papers indicated that the expert was the Teacher, with the next most frequent being an Expert (13.13%). Importantly, in some articles we had to infer who the expert was as this crucial aspect not was explicitly stated. Next, we present some examples. Akkus et al., (2017) indicated that the data was collected by the researcher, but at other times they mentioned “teacher evaluation”. In many of the publications conducted in the medical education context, the authors referred to "Faculty" offering no further information about who were those experts. Evans et al., (2007) referred to the “assessor”, as the agent who performed the expert measurement. Grant et al. (2017) referred to the "preceptor". These were just a few as to visualize the diversity and, in some case, the lack of specificity of some of the publications. In relation to the purpose of self-assessment use in the study, this factor also showed great diversity. While most of the included studies (96.88%) directly or indirectly analysed the accuracy of the self-assessment, there were many studies (43.75%) that additionally analysed factors that could affect the accuracy of the self-assessment or other educational aspects.

[Insert Table 2 near here]

RQ2. What is the average students' self-assessment accuracy?

Of the 160 papers included in this review, 23 did not provide sufficient information to compute the effect size on self-assessment accuracy. 279 standardized mean differences scores were part of the g_p meta-analysis. The first analysis on the total included data showed an average mean effect of g_p = 0.0084, 95%CI [-0.660, 0.677] and not significant, z = 0.024, p = .980. In the case of the meta-analysis of correlations,167 Fisher's z were included. Studies with a sample size less than 4 were excluded from the analysis. The average effect was Fisher's z = 0.501, 95%CI [0.430, 0.572], and statistically significant, z = 13.801, p < .001.

The possible presence of outlier in both meta-analyses was analysed through the analysis of Studentized residua and Cook's distance. Five outliers were identified for g_p (with a g = -27.98, g = -11.28, g = 10.18, g =14.26, and g = 17.51), and two for Fisher's z (Fisher's z = 1.291 and 1.293).

After removing these outliers, we performed the meta-analytical treatment again, this time giving the following results average mean effect was g_p = 0.206, 95%CI [0.061, 0.351] and this time statistically significant, z = 2.795, p = .005. The level of heterogeneity was also large and significant, I^2 = 97.00%, Q(274) = 8,483.644, p < .001. This effect suggested an average effect of slight overestimation of self-evaluations with respect to expert evaluations, indicating a high heterogeneity in the effect found. In the case of the correlations, the average effect after removing the outliers was, Fisher's z = 0.472, 95%CI [0.411, 0.533], and statistically significant, z = 15.09, p < .001, and heterogeneity significant, I^2 = 85.26%, Q(163) = 939.077, p < .001. This effect indicated a significant medium/high correlation between the scores issued by the

students through their self-assessments and the experts' scores, with also a wide heterogeneity in the effect.

Figure 2 shows the funnel plot for both meta-analyses (g_p in panel A and Fisher's z in panel B). The red line in each funnel plot represents the linear regression for the values of each meta-analysis (Egger's test). In the case of g_p the regression is significant, $b_1 = 1.793$, $z = 2.986$, $p = .002$, but it is not for Fisher's z, $b_1 = 0.658$, $z = 1.858$, $p = .063$. As we previously indicated, the analysis of the distributions of the g's presented some asymmetry (although not in the case of the correlations). To further explore the existence of publication bias, we analyzed the p-value distributions using a *p*-curve analysis.

[Insert Figure 3 near here]

This method analyses the possible anomalous distribution of significant *p*-values (Simonsohn et al., 2014). To continue with the analytical logic followed so far, we analysed separately the distribution of studies using contrast of means (*t*-student), and those employing correlations (Figure 3, Panel A for *t's* y panel B for *r's*). The right-skewed of the *p*-curve was significant in both cases, $z = -45.13$, $p < .001$, and $z = -57.49$, $p < .001$ (full t and r *p*-curve respectively), and $z = -45.37$, $p < .001$, and $z = -55.49$, $p < .001$ (half t and r *p*-curve respectively). This distribution shows the veracity of the effect found in the studies. In addition, the *p*-curve did not indicate evidential inadequacy (i.e., flatter than 33% power), $z = 32.21$, $p > .999$, and $z = 43.88$, $p > .999$ (full *t* and *r p*-curve respectively), and $z = 39.97$, $p > .999$, and $z = 47.27$, $p > .999$ (half *t* and *r p*-curve respectively). The estimated power of tests included in the *p*-curve was 99% in both cases. The results of the *p*-curve analysis show that the distribution of the probabilities of the *p*-values in the significant effects showed no evidence of publication bias and/or

p-hacking in the included studies for both analytical treatments (mean and correlation contrasts).

[Insert Figure 4 near here]

RQ3. Do the moderators influence students' self-assessment accuracy?

Tables 3 and 4 summarise the results for the moderation analysis of the proposed factors, both for g_p and Fisher's *z*.

[Insert Table 3 near here]

Regarding the meta-analysis based on contrast of means four factors showed a moderating role on the effect of accuracy on self-assessment (Table 3). These were, having previous experience (SA experience) performing self-assessment ($Q_M(1) = 5.789$, $p = .016$), receiving Feedback ($Q_M(1) = 4.242$, $p = .039$), Content knowledge ($Q_M(1) = 6.135$, $p = .013$), and analysing the Formative assessment role of self-assessment ($Q_M(1) = 100.579$, $p < .001$). It is worth noting that most of the effects calculated showed a tendency to overestimate against the expert's assessment, and in most cases, the effect found for each moderator level is significantly different from 0. It is particularly worth noting the moderating role of Feedback. We found that when self-assessment was performed in the absence of Feedback, a significant overestimation effect was found ($g = 0.267$, $p > .001$), but when offered Feedback the self-assessment showed a near perfect fit ($g = 0.083$, $p = .637$). Surprisingly, the Use of Rubric for self-assessment did not show a significant improvement in the accuracy of such assessments, nor did knowing the Assessment criteria to be used in the assessment. A striking result was that the studies that investigated the Formative assessment role of using self-assessment this showed a moderating effect on accuracy, in that the participants on those studies overestimated ($g = 0.415$, $p <.001$).

Regarding the results of the meta-analysis based on correlations, the only two factors that showed a significant moderating role were Feedback ($Q_M(1)= 6.264$, $p = .012$) and Educational level ($Q_M(3)= 10.274$, $p = .016$).

[Insert Table 4 near here]

Again, it should be noted that few effects showed an effect size greater than 0.5 (a value to start considering large effects, Cohen, 1988), even though most of the levels of each moderator showed a significant correlation effect ($p < .05$). Similar to the other meta-analysis, facilitating Feedback improved accuracy on self-assessment ($z = 0.559$, $p > .001$). Additionally and unlike what was seen in g_p, no moderation effects were found for Context knowledge, nor Formative assessment, although SA experience was close to significance ($Q_M(1) = 3.308$, $p = .068$). The moderator Field of knowledge yields high correlations but, in some of the areas, there were small sample sizes (i.e., $z = 0.867$, $p = .001$, $k = 1$, for Mathematics). Also surprising are the high correlations found for the Primary education ($z = 0.791$, $p < .001$, $k = 5$) and Secondary education ($z = 0.717$, $p < .001$) for Educational level.

[Figure 5 near here]

Finally, we analysed the extent to which the presence of the moderators (e.g., those studies in which the moderator assessed was shown to be present and coded as YES) could predict self-assessment accuracy. Figure 5 shows the effect that the presence of the moderating factors had on self-assessment accuracy in the case of mean differences (in the Panel A) and correlations (in the Panel B). A (mixed-effects) multilevel meta-regression model was fit for *g* and for Fisher's correlations, including the proposed factor as a source of precision on self-assessment as a predictor/moderator.

In the case of *g* the analysis showed a significant difference as a function of the moderating variable $Q_M(6) = 63.481$, $p < .001$. The isolated analysis of each subgroup

of moderators showed that know the Assessment criteria (g = 0.142, p = .102), SA experience (g = 0.109, p = .352), Feedback (g = 0.016, p = .868), y Content knowledge (g = 0.085, p = .420), showed a non-significant effect, indicating that the studies in which these moderators were investigated did not differ from 0 and, therefore, these self-assessment were more accurate. In contrast, the difference between the types of moderators was not significant in the studies with correlations, $Q_M(6) = 1.2519$, p = .974. This means that all correlation effects associated with the presence of all moderators were significant ($p < .05$). However, no significant differences were found when comparing the effects of each biter (none was significantly greater or less significant than the rest).

Discussion

Our aim was to investigate the empirical evidence around the students' self-assessment accuracy and its moderators' effects. We explored three research questions which results will be discussed next.

RQ1. Characteristics of self-assessment accuracy studies

An interesting first finding is that self-assessment scoring accuracy is a topic that generates considerable interest as shown by the fact that, regardless of our strict inclusion criteria, there were a high number of studies included in this meta-analysis. This result is in line with previous reviews that have shown self-assessment scoring accuracy to be the prevalent line of research in self-assessment (e.g., Andrade, 2018, 2019). Another interesting finding is the uneven distribution of explored knowledge areas, with some receiving minimal exploration (e.g. Nursing, Computer or Biology) while others have been largely explored such as medicine. This situation was already identified by Falchikov and Boud (1989) in the first meta-analysis of the field, so researchers still need to amplify research in some knowledge areas.

As expected, most of the research has been conducted in higher education, a finding shared with other reviews in the assessment field (e.g., Panadero & Alqassab, 2019). While there are different reasons for this, it would be crucial that scholars increase the study of self-assessment accuracy in other educational levels as previous studies have found different effects (Sánchez et al., 2017).

When it comes to the characteristics of the implementation of the self-assessment itself, first, in more than half of the studies the participants were given the assessment criteria in different forms (e.g., out loud), with the most salient one being rubrics used in around half of the studies. Second, in above three quarters of the studies the expert used for the calculation of the accuracy was the teacher. Third, in almost all, the main objective was a direct or indirect analysis of the accuracy of self-assessment; interestingly 40% also explored factors that could influence accuracy such as gender, formative value of evaluation, use of digital media, or different forms of feedback, for example. There were other aspects that were explored such as the experience of the students performing self-assessment or whether they received feedback on their task performance or not. This leads to a crucial conclusion of our study, there large heterogeneity in the interventions and variables studied. This has two sides. On the one hand, it is excellent news that self-assessment accuracy is studied across different types of interventions and contexts (especially in higher education) as its effects are tested in many different and complex situations. On the other hand, that makes it more difficult to extract strong conclusions as the comparison is not as straightforward, especially with the low quality in the report of the studies. In this sense, a large proportion of the studies missed crucial information (e.g., who was the expert, steps of the procedure). This is something found in previous assessment research reviews (e.g., Ferrero et al., 2021;

Panadero & Alqassab, 2019; Panadero et al., 2017), that researchers in the field should try to correct.

What are the implications of our findings from RQ1 for the self-assessment and formative assessment fields? We believe three are of particular significance. First, given the extensive study of scoring accuracy, it may be important for formative scholars to consider integrating its educational value into their self-assessment models and concepts. Second, the high volume of published research cannot be properly interpreted and integrated unless primary studies provide accurate reports on methodological and intervention characteristics. Third, researchers should not limit their exploration to scoring accuracy alone but should also investigate content accuracy, both independently and in relation to scoring accuracy.**RQ2. Average students' self-assessment accuracy**

Before presenting the empirical results, we will discuss a crucial feature of our review. We computed two different meta-analyses based in the type of statistic used in the original studies, either contrast of means or correlations. To compute the effects of the contrast of means studies we used standardized mean different test scores and to compute the effects of the correlations studies we used Fisher's *Z* scores. This is interesting in itself because it allowed us to (a) investigate the specific average effect for each group of studies and (b) explore the sensibility of each method. While correlational studies only indicated the level of relationship among the students' self-awarded score and the one by the expert, the method of contrast of means allowed us to explore in greater detail aspects such as the type of bias or the direction of the inaccuracy (over estimation vs. under estimation). In that sense, contrast of means was a more interesting method. Additionally, this have been further contrasted through our moderators, which will be discussed in the next section.

Regarding the results, after eliminating the outliers, both meta-analyses showed significant results with mean effects of $g_p = 0.206$ and $z = 0.472$. A $g = 0.206$ indicates the size of the overestimation of the students when they score their own work as in this statistic values closer to 0 indicate better accuracy. On the other hand, a $z = 0.472$ reports a positive correlation of intermediate size, which in this statistic should be interpreted as that the students' self-assessment score and the expert's score, usually the teacher, share an important part of variance as here values closer to 1 indicate perfect overlap. Therefore, our results show that, while there is a significant portion of shared variance among students' and teachers' scores, the direction of deviation clearly leans towards overestimation. We now compare our results to the two previous meta-analyses that have investigated the same construct.

We start by comparing the meta-analytic methods. Falchikov and Boud (1989) calculated, based on the information provided by the included studies, Cohen's *d*, *r* correlations coefficient*,* and proportions of agreement. Brown and Harris (2013) calculated Cohen's *d* effect size. As just explained, in our study we calculated a Hedges' *g* to obtain the effect sizes associated with the contrast of means. Although this method is very similar to *d*, it proposes a bias correction in the case of small samples (Borenstein et al., 2009). In the case where the correlation coefficient could be calculated, these were transformed to Fisher's *z* scores in order to calculate the variance of the effect. In some ways, our calculation of the effect size for the study is more accurate and sensitive than in previous studies. More importantly, in our work, before analyzing the average effects, as well as the moderation effects, we analyze the possible existence of outliers. Failure to perform this treatment can lead to the inclusion of potentially unrealistic effect sizes that distort the final result (Ferrero et al., 2021; Viechtbauer & Cheung, 2010). In addition, we then analyzed whether the set of results

analyzed showed any type of publication bias (both through the distribution of effect sizes and the distribution of p-values).

Now, comparing our results to those of the two previous meta-analyses, there are clear alignments. First, the three meta-analysis found an overlap between students' self-assessment scoring and that of the teacher as found in our results, in Falchikov and Boud (1989) who reported a 64.1% agreement, and in Brown and Harris (2013) who found correlations within a wide range ($r \approx 0.20$ to 0.80) -though only a limited number of studies documented correlations exceeding 0.60. Therefore, it seems like there is a clear connection between the score that students give to themselves when compared with an expert, usually the teacher. Secondly, in terms of the direction of the deviation, our results align with the tendency for overestimation found by Falchikov and Boud (1989) mean effect size of 0.47. Importantly, we only integrate our results with those of the two previous meta-analyses because it would not make sense to integrate it with the primary studies that our own meta-analysis has just analysed.

What are the implications of our findings from RQ2 for the self-assessment and formative assessment fields? These are two important findings. In one hand there is a significant overlap between how students grade themselves when compare with teachers. In our view this could give us some reassurance that teachers and students are "speaking similar languages" when it comes to grading. In the other hand, the general tendency towards overestimation indicates that self-grading is not an indicator that we can take as face value. It is known from Dunning – Krueger effect that humans tend to hold overly favourable views on their capabilities (Kruger & Dunning, 1999). They also found that, it was possible to correct for such miscalibration by exposing the individuals to practice and observation of others' performance, among others. Thus, one conclusion that we can extract is that it is logical that students tend to overestimate due to their

limited knowledge on the task at hands (Panadero et al., 2016). When we consider that most of the studies are conducted in higher education, academic performance at such level is based on specific aspects of the particular subject. Being that students change subjects frequently, once they start a new one, they can be considered novices and, therefore, incapable of producing an accurate self-assessment. Thus, the need to build up on the positive news, the fact that there is a considerable overlap between teachers' and students' scores and reflect about the need to let students get to know the subject while implementing interventions (e.g., rubric, feedback) that enhance their self-assessment skill. Thus, the need for the exploration of moderators as to try to disentangle their effects.

RQ3. Moderators and their effects

Two important aspects when considering our analyses of the moderators are the high heterogeneity of self-assessment interventions and the low level of detail about the research methods and intervention characteristics showed by a large percentage of the included studies. We analysed the effects of nine moderators using both types of meta-analyses, contrast of means and correlations. It is worth reminding that most of the moderators were coded dichotomously (Yes or No) for three reasons: (a) our study had an explanatory nature including as many moderators as possible, (b) a more advanced coding of the categories would have extended quite significantly the length of the manuscript, and (c) in a significant number of the included studies there were missing information as to create more elaborated coding. In general, considering the *g* values, the general trend of the effects for each level of the moderators analyzed was overestimation, in greater or lesser extent depending on the moderator; while considering the correlations, most of them were non-significant. We will next, discuss

the moderators individually and point out the ones with results explored in the two previous meta-analyses.

Assessment criteria. This factor was not a significant moderator of the effect in either meta-analysis calculations. However, although facilitating or not the evaluation criteria did not show differences in the *g* analysis, facilitating the criteria showed a non-significant effect in the case of the *g's* ($g = 0.142$, $p = .103$). This would indicate that the average effect of the studies where this factor is met the differences between self-evaluations and expert evaluations was not significant. It was to be expected that students that were given clear criteria would be able to be more accurate (e.g., Kitsantas & Zimmerman, 2006), as also discussed briefly in Brown and Harris (2013) meta-analysis. A crucial aspect of self-assessment are the internal criteria and standards established by the students (Butler & Winne, 1995). In academic settings, these need to be contrasted with the external criteria and standards given by the teachers as to correct for bias and increase accuracy (Nicol, 2021), in what has been called "making the implicit explicit to correct for self-bias" (Panadero et al., 2019). That is why, providing assessment criteria when asking students to self-assess has been widely recommended.

Use of rubric. This factor did not show a significant moderating effect in both meta-analyses. In the case of the analysis through the *g's*, facilitating the rubric for self-evaluation showed a non-significant effect ($g = 0.185$, $p = .068$) and a lower effect than not using it ($g = 0.231$, $p = .034$). Regarding correlations, these were higher when rubrics were used ($z = 0.515$) than when rubrics were not used ($z = 0.416$). Therefore, our results indicate that rubrics, while not significantly, decreased the deviation as measured by *g* and enhance the overlap with expert's scores as measured by the correlations. This would be in alignment with two of the conclusions from the review conducted by Jonsson and Svingby (2007). First, as discussed by those authors, rubrics

can enhance the reliability of the scoring of performance assessments. However, they also explored this in combination with exemplars use, rater training and when some rubrics characteristics were present (i.e., analytic and topic-specific). To the dichotomous nature of our coding, these aspects were not explored here and that also diluted the possibilities to differentiate between strong and weak rubrics studies. Therefore, second, our results in a way also aligned with a second conclusion from Jonsson and Svingby, that is rubrics do not facilitate valid judgment of performance assessments per se. Many aspects influence the impact of rubrics on students (Jonsson & Svingby, 2007; Panadero & Jonsson, 2013, 2020) and our study did not go into such details as our aim was more general that the one from Jonsson and Svingby who just focused on rubrics. Brown and Harris (2013) also discussed the enabling power of rubrics to enhance accuracy.

Self-assessment experience. Experience is one of the factors that modulates self-assessment accuracy of self-assessments. While it did not fully correct for the overestimation as the effect size was different from zero in a positive direction in the experience group, it improved significantly the accuracy (g = 0.177) and a non-significant better overlap between students' and teachers' scores (z = 0.512). The positive effect on accuracy of having experience in self-assessment have been greatly discussed in previous research (e.g., Panadero et al., 2016; Sluijsmans et al. 2002). Importantly, the two previous meta-analyses did not explore specifically the role of previous self-assessment experience. However, Falchikov and Boud (1989) mentioned “Self-assessment may be regarded as a skill and, as such, needs to be developed” (p. 426), while Brown and Harris (2013) also mentioned that “Training is also likely to improve accuracy” (p. 386). From this moderator, we can conclude that students become more accurate in self-assessment with increased practice. It is crucial to note

that this practice may also be linked to students' content knowledge, which was another significant moderator. Hence, these two factors should be recognized as closely interconnected.

Feedback. This was the only moderator significant in both meta-analyses calculations. Providing feedback to the students, mostly on their academic performance, improved the closeness of students' and teachers' scores as the studies with feedback showed effect sizes closer to 0 and higher correlations (Figure 5). Also, the correlations were significantly higher showing greater overlap among students' and teachers' scores. This result is attested by previous research (e.g., Baars et al., 2014; Dunlosky et al., 2011; Lipko et al., 2009; Sitzmann et al., 2010) and the recommendations have been for a long time that, to improve the effects of self-assessment, students need to receive external feedback to contrast the veridicality of their judgments (e.g., Andrade & Valtcheva, 2009; Eva & Regehr, 2005, 2008). In the theoretical framework, we presented that previous research had shown contradictory results. It is our hope that, as meta-analytical reviews integrate existing primary research, our results can shed light on the importance of feedback in self-assessment, a topic greatly discussed in previous research (Boud, 1995). Therefore, from an educational perspective, it is clear that teachers should plan to give feedback when they ask their students to self-assess.

Content knowledge. Having knowledge or mastery about the topic to be self-assessed showed a significantly increased precision though it did not reach perfect precision either. However, the correlation results did not show significant differences. Previous research has emphasized that having expertise in the task at hands might be crucial for accurate self-assessment (e.g., Brown & Harris, 2013; Dochy et al., 1999; Panadero et al., 2016). This is logical as students need to have precise criteria in the particular task and knowledge about the standards, in other words, how a good performance of the task

looks like. In general, previous self-assessment research has focused in self-assessment previous experience or content knowledge. It might be important, in light of our results, to consider both as important and working in integration.

Incentive. This moderator was not significant in either of the meta-analyses. It could have been expected that, if students receive an incentive for being accurate, this might increase their accuracy. Actually, it was one of the points argued in one of the previous meta-analysis: "Systems where students predicted or monitored their accuracy and achievement and/or rewarded themselves for accuracy or improvement also were correlated with gains." (Brown & Harris, 2013 p. 381). However, there are two important limitations regarding the calculation of this moderator effects: it is the less studied (only 23 and 19 publications included incentives), and the reports did not always clarify what was the intention of the incentive. Although the incentives were aimed at trying to motivate students to be accurate in their self-assessments, in very few cases students had information about what would be considered accurate, and this could somewhat mislead students into thinking that the incentive would be associated with participation.

Formative assessment. This moderator showed a significant effect only in the *g's*. Somewhat surprisingly, in studies where the formative role of self-assessment was emphasized students tended to show a significantly larger overestimation bias (g = 0.415, $p < .001$), than when this factor was not measured (g = 0.174, p = .016). While this might seem like a counterintuitive result, it is actually logical and we next propose an explanation for it. Some formative scholars have doubts about the validity and importance of scoring accuracy (discussed in Andrade, 2018, 2019), and disregard the use of self-assessment with summative purposes. When the self-assessment intervention has formative basis, as it was the case in 37 and 22 studies -corresponding with our two

meta-analysis calculations-, the intervention focus is more on practicing self-assessment, promoting reflection or enhancing learning, rather than emphasizing scoring accuracy (e.g., McKevitt, 2016; Nederhand, 2000; Pilotti et al., 2019). The researcher, the teacher (which in some of the included studies is the researcher), or the students might have experienced that, for formative purposes, scoring accuracy is irrelevant. In future research, it would be interesting to explore whether, while keeping the emphasis in the formative purposes, students would also receive the message that scoring accuracy is important if they anchor their score to the learning goals, something that we have not found in the included studies. This line of research would be particularly fruitful as to understand the complexity of implementing self-assessment in the classroom, especially in higher education contexts, where summative and formative purposes coexist.

Field of knowledge. While this factor was not a significant moderator, a close analysis offers interesting information to consider. Firstly, there is an unbalanced distribution across the areas. While some show a great interest (with medicine in the lead $k = 84/68$) there are others with low or null presence. For this reason, effect sizes should be interpreted with caution in areas where the sample size are low. Regarding the effects found, it is worth highlighting the one obtained in education ($g = -0.141$, $p = .507$, and $z = 0.555$, $p < .001$) which, although it shows a negative average effect, the absence of significance indicates that this value does not differ from 0 indicating good accuracy in the self-assessments. Similarly as in Medicine, where accuracy is very good through the g's ($g = 0.006$, $p = .967$) although not so good through the correlations ($z = 0.423$, $p < .001$). In the opposite example we find areas of knowledge that show large biases in precision, for example Chemistry with a large overestimation ($g = 0.965$), although with large variability of the results and few cases ($k = 8$), or the area of science with a large

underestimation (g = -0.625, k = 3). Importantly, there are very few studies comparing self-assessment in different fields of knowledge (Panadero et al., 2020), which is the type of research we need to better understanding this effect. If we compared our results to those of Falchikov and Boud (1989) they found that "science" subjects produced more accurate self-assessments than social science. This could be hypothesized as a result of the different tasks nature, as some of the activities in science subjects might produce more unequivocal results. However, this is to be tested as, to our knowledge there is no previous research comparing the scoring accuracy in different subjects. While doing it, the researchers conducting such study should keep in mind that Falchikov and Boud (1989) also found that the nature of the task did not influence accuracy, though their categories were three that look unrelated to science vs. social science: professional practice, traditional academic activity focus on process, and traditional academic activity focus on product.

Educational level. This factor showed a significant effect in the case of the correlations and close to this through the g's. If we look at the correlations, the studies performed in primary and secondary education showed very high correlation effects (z = 0.791 and z = 0.717 respectively), on the contrary the correlations in undergraduate and graduate studies showed low correlations, being non-significant in the case of graduate studies (z = 0.440 and z = 0.382 respectively). Note that the number of studies is very unbalanced among educational levels with higher education being the most studied. If we look at the effects through the g's the pattern seen through the correlations is repeated, with the Primary and Secondary stages showing better accuracy than the University and Postgraduate levels. Something interesting is the significant underestimation effect found in Graduate studies (g = -0.750). These results somehow support the already reported relationship (see León et al., 2021) between proficiency level and

overestimation and underestimation biases. We hypothesize that there might be at least two reasons for this result. First, primary and secondary education students might have less at stake because of their grades, and therefore they would be less pressured to overestimate. Second, university and postgraduate students perform more complex and multifaceted tasks, ergo more difficulty in reaching accuracy. As with field of knowledge, we need research that directly explores differences among educational levels. Importantly, the two previous meta-analyses did not explore such educational level differences because they focused on specific levels. It is also relevant to point out that there is barely no empirical studies comparing in the same publication different levels except for a few exceptions (e.g., Panadero et al., 2022; Yan, 2018).

Limitations

First, our coding of the moderators characteristics was mostly dichotomous due to, both, the high number of moderators included and the descriptive nature of this meta-analysis. Nevertheless, these moderators are multifaceted and offer more complex learning effect (e.g., who created the rubric? Was it a valid and reliable instrument? In which moment did the students know about the assessment criteria?) Therefore, while it is not feasible to run a meta-analysis with nine moderators including such multifaceted information, this is still a limitation in our interpretation of the data. Nevertheless, it is important to remark that our study has an explanatory nature including as many moderators as possible to open venues for future, more concrete, meta-analysis on this topic.

Second, there are limitations based in the quality of the academic studies included. As mentioned earlier, not all the studies have strong research design, control of strange variables, etc. And third, yet related to the previous one, a significant number of studies have problems in their reporting of methods and results. It is not unusual that

crucial pieces of information are missing (e.g., Ferrero et al., 2021; Panadero & Alqassab, 2019; Panadero et al., 2017) and some solutions are being proposed (Panadero et al., 2023). These two limitations had a very significant impact in the opportunity to perform deeper analyses.

Future lines of research

Next, we propose future lines of research. First, it would be important for the field to explore the effects of more students' characteristics in the scoring accuracy as the general field of feedback is moving towards paying greater attention to individual differences (Panadero, 2023). For example, it would be of interest to strengthen our research on individual variables such as: academic achievement level, personality traits, self-regulatory skills, or gender. While two have been explored here –i.e., self-assessment experience and content knowledge- these are more instructional and closer to the task than the ones we are proposing.

Second, it would be relevant to explore in more detail the moderators explored here because, as mentioned earlier, there is much more information that if coded specifically would allow for a greater understanding of these moderators effects.

And third, it would be important to conduct studies not only around scoring accuracy but on content accuracy (Panadero et al., 2016). As important as it can be that students are able to calculate their score, it is probably even more important if students are able to identify what is wrong and right in their performance as to make pertinent adjustments, which at the end, will affect their score. Unfortunately, there are no content accuracy studies to our knowledge.

Conclusion

We investigated students' self-assessment scoring accuracy and its moderators. We believe ours was an important enterprise for two reasons. First, to help clarify what

are the main results on the most prolific area of research in self-assessment (Andrade, 2018, 2019). And second, there might be a learning gain in asking students to accurately estimate their grade when this is anchored to appropriate criteria, good examples, etc. (Panadero et al., 2016). At the end, self-assessment when done reflectively should make our students evoke their performance, compare to their standards and reach an evaluation of the quality of their work. All of which should benefit students' learning.

An important characteristic of our work is that we used two methods to calculate accuracy: contrast of means and correlations. Our results show that students tend to overestimate, but they do less so when they receive feedback, have larger self-assessment experience and content knowledge, when there are no formative purposes, and when younger students are involved (primary and secondary education). Nevertheless, it is imperative that we continue developing the field by (a) designing stronger studies, (b) improving dramatically the quality of the reports, (c) performing replication studies, and (d) approaching self-assessment accuracy from more complex angles (Panadero et al., 2016).

References

* (Marked with an * the studies included in the review).

*Abadel, F. T., & Hattab, A. S. (2013). How does the medical graduates' self-assessment of their clinical competency differ from experts' assessment? *BMC medical education, 13*(1), 1-9. https://doi.org/10.1186/1472-6920-13-24

*Abate, M. A., & Blommel, M. L. (2007). Self-assessment tool for drug information advanced pharmacy practice experience. *American Journal of Pharmaceutical Education, 71*(1): 02. https://doi.org/10.5688%2Faj710102

*Abdalla, R., Bishop, S. S., Villasante-Tezanos, A. G., & Bertoli, E. (2021). Comparison between students' self-assessment, and visual and digital assessment techniques in dental anatomy wax-up grading. *European Journal of Dental Education, 25*(3), 524-535. https://doi.org/10.1111/eje.12628

*Abeyaratne, C., Nhu, T., & Malone, D. (2022). Self-Assessment of Therapeutic Decision-Making Skills in Pharmacy Students. *American Journal of Pharmaceutical Education*, *86*(4). Retrieved from https://www.ajpe.org/content/ajpe/86/4/8696.full.pdf

*Admiraal, W., Huisman, B., & Pilli, O. (2015). Assessment in massive open online courses. *Electronic Journal of E-learning, 13*(4), pp207-216. Retrieved from https://www.academic-publishing.org/index.php/ejel/article/view/1728/1691

*Agrawal, S., Norman, G. R., & Eva, K. W. (2012). Influences on medical students' self-regulated learning after test completion. *Medical education, 46*(3), 326-335. https://doi.org/10.1111/j.1365-2923.2011.04150.x

*Aiko, M. (2018). The Relationships Between the Accuracy of Self-Evaluation, Kanji Proficiency and the Learning Environment for Adolescent Japanese Heritage Language Learners. *Journal of Language and Education, 4*(2), 6-23. https://doi.org/10.17323/2411-7390-2018-4-2-6-23

*Aitken, A., & Thompson, D. G. (2018). Using software to engage design students in academic writing. *International Journal of Technology and Design Education, 28*(3), 885-898. https://doi.org/10.1007/s10798-017-9413-4

*Akkuş, H., & Sinem, Ü. N. E. R. (2017). The effect of microteaching on pre-service chemistry teachers' teaching experiences. *Cukurova University Faculty of Education Journal, 46*(1), 202-230. https://doi.org/10.14812/cuefd.309459

*Akyuz, D. (2018). Measuring technological pedagogical content knowledge (TPACK) through performance assessment. *Computers & Education, 125*, 212-225. https://doi.org/10.1016/j.compedu.2018.06.012

*Alameddine, M. B., Englesbe, M. J., & Waits, S. A. (2018). A video-based coaching intervention to improve surgical skill in fourth-year medical students. *Journal of surgical education, 75*(6), 1475-1479. https://doi.org/10.1016/j.jsurg.2018.04.003

*Albanese, M., Dottl, S., Mejicano, G., Zakowski, L., Seibert, C., Van Eyck, S., & Prucha, C. (2006). Distorted perceptions of competence and incompetence are more than regression effects. *Advances in health sciences education, 11*(3), 267-278. https://doi.org/10.1007/s10459-005-2400-7

*Alfakhry, G., Mustafa, K., Alagha, M. A., Milly, H., Dashash, M., & Jamous, I. (2022). Bridging the gap between self-assessment and faculty assessment of clinical

performance in restorative dentistry: A prospective pilot study. *Clinical and Experimental Dental Research*, *8*(4), 883-892. https://doi.org/10.1002/cre2.567

*Ammentorp, J., Thomsen, J. L., Jarbøl, D. E., Holst, R., Øvrehus, A. L. H., & Kofoed, P. E. (2013). Comparison of the medical students' perceived self-efficacy and the evaluation of the observers and patients. *BMC Medical Education, 13*(1), 1-6. https://doi.org/10.1186/1472-6920-13-49

*Andoh, B., & Jones, P. (2008). Students' self-assessment in law. *The Law Teacher, 42*(2), 200-212. http://dx.doi.org/10.1080/03069400.2008.9959776

Andrade, H. (2018). Feedback in the context of self-assessment. In A. A. Lipnevich & J. K. Smith (Eds.), *The Cambridge handbook of instructional feedback* (pp. 376–408). Cambridge University Press.

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. https://doi.org/10.7275/g367-ye94

Andrade, H., & Valtcheva, A. (2009). Promoting learning and achievement through self-assessment. *Theory Into Practice, 48*(1), 12-19. https://doi.org/10.1080/00405840802577544

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA publications and communications board task force report. *American Psychologist, 73,* 3–25. https://psycnet.apa.org/doi/10.1037/amp0000389

*Aryadoust, V. (2015). Self-and peer assessments of oral presentations by first-year university students. *Educational Assessment, 20*(3), 199-225. https://doi.org/10.1080/10627197.2015.1061989

*Ashton, K. (2014). Using self-assessment to compare learners' reading proficiency in a multilingual assessment framework. *System, 42*, 105-119. http://dx.doi.org/10.1016/j.system.2013.11.006

*Austin, Z., & Gregory, P. A. (2007). Evaluating the accuracy of pharmacy students' self-assessment skills. *American journal of pharmaceutical education, 71*(5). https://doi.org/10.5688%2Faj710589

*Baars, M., Vink, S., van Gog, T., de Bruin, A., & Paas, F. (2014). Effects of training self-assessment and using assessment standards on retrospective and prospective monitoring of problem solving. *Learning and Instruction, 33*, 92-107. https://doi.org/10.1016/j.learninstruc.2014.04.004

*Baecher, L., Kung, S. C., Jewkes, A. M., & Rosalia, C. (2013). The role of video for self-evaluation in early field experiences. *Teaching and Teacher Education, 36*, 189-197. https://doi.org/10.1016/j.tate.2013.08.001

*Balch, W. R. (1992). Effect of class standing on students' predictions of their final exam scores. *Teaching of Psychology, 19*(3), 136-141. https://doi.org/10.1207/s15328023top1903_1

*Baleghizadeh, S., & Hajizadeh, T. (2014). Self-and Teacher-Assessment in an EFL Writing Class. *Gist: Education and Learning Research Journal,* (8), 99-117. Retrieved from https://dialnet.unirioja.es/descarga/articulo/4774787.pdf

*Ballantine, J. A., Larres, P. M., & Oyelere, P. (2007). Computer usage and the validity of self-assessed computer competence among first-year business students. *Computers & Education, 49*(4), 976-990. https://doi.org/10.1016/j.compedu.2005.12.001

*Bergee, M. J. (1993). A comparison of faculty, peer, and self-evaluation of applied brass jury performances. *Journal of Research in Music Education, 41*(1), 19-27. https://doi.org/10.2307%2F3345476

*Bergee, M. J. (1997). Relationships among faculty, peer, and self-evaluations of applied performances. *Journal of Research in Music Education, 45*(4), 601-612. https://doi.org/10.2307%2F3345425

*Biango-Daniels, M., & Sarvary, M. (2021). A challenge in teaching scientific communication: academic experience does not improve undergraduates' ability to assess their or their peers' writing. *Assessment & Evaluation in Higher Education, 46*(5), 809-820. https://doi.org/10.1080/02602938.2020.1812512

*Biernat, K., Simpson, D., Duthie, E., Bragg, D., & London, R. (2003). Primary care residents self assessment skills in dementia. *Advances in health sciences education, 8*(2), 105-110. https://doi.org/10.1023/A:1024961618669

*Biswas, S. S., Jain, V., Agrawal, V., & Bindra, M. (2015). Small group learning: effect on item analysis and accuracy of self-assessment of medical students. *Education for health, 28*(1), 16. https://doi.org/10.4103/1357-6283.161836

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice, 5*(1), 7–73. https://10.1080/0969595980050102

*Boerebach, B. C., Arah, O. A., Busch, O. R., & Lombarts, K. M. (2012). Reliable and valid tools for measuring surgeons' teaching performance: Residents' vs. self evaluation. *Journal of surgical education, 69*(4), 511-520. https://doi.org/10.1016/j.jsurg.2012.04.003

*Bolívar-Cruz, A., Verano-Tacoronte, D., & Galván-Sánchez, I. (2018). Do self-efficacy, incentives and confidence in public speaking influence how students self-assess?/¿ Influyen la autoeficacia, los incentivos y la confianza para hablar en público en cómo se autoevalúan los estudiantes?. *Cultura y Educacion, 30*(3), 528-555. https://doi.org/10.1080/11356405.2018.1488420

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta–analysis.* Chichester, United Kingdom: John Wiley & Sons. https://doi.org/10.1002/9780470743386

Boud, D. (1995). *Enhancing learning through self-assessment*. New York: RoutledgeFalmer.

*Boud, D., Lawson, R., & Thompson, D. G. (2013). Does student engagement in self-assessment calibrate their judgement over time?. *Assessment & Evaluation in Higher Education, 38*(8), 941-956. https://doi.org/10.1080/02602938.2013.769198

*Boud, D., Lawson, R., & Thompson, D. G. (2015). The calibration of student judgement through self-assessment: disruptive effects of assessment patterns. *Higher education research & development, 34*(1), 45-59. https://doi.org/10.1080/07294360.2014.934328

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

Brown, G. T. L., & Harris, L. R. (2013). Student self-assessment. In J. H. McMillan (Ed.), *The SAGE handbook of research on classroom assessment* (pp. 367–393). Thousand Oaks: Sage.

Butler, D. L. (2011). Investigating Self-Regulated Learning Using In-Depth Case Studies: University of British Columbia, Vancouver, Canada. In *Handbook of self-regulation of learning and performance* (pp. 360-374). Routledge.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: a theoretical synthesis. *Review of Educational Research, 65*(3), 245–281. https://doi.org/10.3102%2F00346543065003245

*Butler, Y. G., & Lee, J. (2006). On-task versus off-task self-assessments among Korean elementary school students studying English. *The Modern Language Journal, 90*(4), 506-518. https://doi.org/10.1111/j.1540-4781.2006.00463.x

*Butterworth, K. (2010). Developing self-assessment skills amongst doctors in Nepal. *Medical Teacher, 32*(2), e85-e95. https://doi.org/10.3109/01421590903206372

*Carroll, D. (2020). Observations of student accuracy in criteria-based self-assessment. *Assessment & Evaluation in Higher Education, 45*(8), 1088-1105. https://doi.org/10.1080/02602938.2020.1727411

Cascallar, E., Boekaerts, M., & Costigan, T. (2006). Assessment in the evaluation of self-regulation as a process. *Educational Psychology Review, 18*(3), 297-306. https://doi.org/10.1007/s10648-006-9023-2

*Cave, J., Washer, P., Sampson, P., Griffin, M., & Noble, L. (2007). Explicitly linking teaching and assessment of communication skills. *Medical Teacher, 29*(4), 317-322. https://doi.org/10.1080/01421590701509654

*Chang, C. C., Liang, C., & Chen, Y. H. (2013). Is learner self-assessment reliable and valid in a Web-based portfolio environment for high school students?. *Computers & Education, 60*(1), 325-334. https://doi.org/10.1016/j.compedu.2012.05.012

*Chur-Hansen, A. (2000). Medical students' essay-writing skills: criteria-based self-and tutor-evaluation and the role of language background. *Medical education, 34*(3), 194-198. https://doi.org/10.1046/j.1365-2923.2000.00457.x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2. Auflage)*. Hillsdale, NJ: Erlbaum.

*Cooney, C. M., Aravind, P., Lifchez, S. D., Hultman, C. S., Weber, R. A., Brooke, S., & Cooney, D. S. (2021). Differences in operative self-assessment between male and female plastic surgery residents: A survey of 8,149 cases. *The American*

Journal of Surgery, *221*(4), 799-803. https://doi.org/10.1016/j.amjsurg.2020.04.009

*Das, M. (1998). Self and tutor evaluations in problem-based learning tutorials: is there a relationship?. *Medical education, 32*(4), 411-418. https://doi.org/10.1046/j.1365-2923.1998.00217.x

*Davey, K. R. (2015). Student self-assessment: Results from a research study in a level IV elective course in an accredited bachelor of chemical engineering. *Education for Chemical Engineers, 10*, 20-32. https://doi.org/10.1016/j.ece.2014.10.001

*De Grez, L., Valcke, M., & Roozen, I. (2012). How effective are self-and peer assessment of oral presentation skills compared with teachers' assessments?. *Active Learning in Higher Education, 13*(2), 129-142. https://doi.org/10.1177%2F1469787412441284

*Dikici, A. (2009). An Application of Digital Portfolio with the Peer, Self and Instructor Assessments in Art Education. *Active Learning in Higher Education, 13*(2), 129-142. https://doi.org/10.1177/1469787412441284

Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer- and co-assessment in higher education. A review. *Studies in Higher Education, 24*(3), 331–350. https://doi.org/10.1080/03075079912331379935

*Domínguez, C., Jaime, A., Sánchez, A., Blanco, J. M., & Heras, J. (2016). A comparative analysis of the consistency and difference among online self-, peer-, external-and instructor-assessments: The competitive effect. *Computers in Human Behavior, 60*, 112-120. https://doi.org/10.1016/j.chb.2016.02.061

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: implications for health, education, and the workplace. *Psychological Science in the Public Interest, 5*(3), 69–106. https://doi.org/10.1111%2Fj.1529-1006.2004.00018.x

*Durant, M. (2014). *Comparison of high school and expert judges' evaluation of high school band soloists* [Doctoral dissertation]. Texas Tech University. http://hdl.handle.net/2346/58515

*Elhadi, M., Ahmed, H., Khaled, A., Almahmoudi, W. K., Atllah, S. S., Elhadi, A., & Esahli, H. (2020). Informed self-assessment versus preceptor evaluation: a comparative study of pediatric procedural skills acquisition of fifth year medical students. *BMC Medical Education, 20*(1), 1-8. https://doi.org/10.1186/s12909-020-02221-2

*Ellery, K., & Sutherland, L. (2004). Involving students in the assessment process. *Perspectives in Education, 22*(1), 99-110. https://hdl.handle.net/10520/EJC87239

*Emam, H. A., Jatana, C. A., Wade, S., & Hamamoto, D. (2018). Dental student self-assessment of a medical history competency developed by Oral and Maxillofacial Surgery Faculty. *European Journal of Dental Education, 22*(1), 9-14. https://doi.org/10.1111/eje.12222

*Ericson, D., Christersson, C., Manogue, M., & Rohlin, M. (1997). Clinical guidelines and self-assessment in dental education. *European Journal of Dental Education*, *1*(3), 123-128. https://doi--org.ujaen.debiblio.com/10.1111/j.1600-0579.1997.tb00021.x

*Esfandiari, R., & Myford, C. M. (2013). Severity differences among self-assessors, peer-assessors, and teacher assessors rating EFL essays. *Assessing writing, 18*(2), 111-131. https://doi.org/10.1016/j.asw.2012.12.002

Eva, K. W., & Regehr, G. (2005). Self-assessment in the health professions: a reformulation and research agenda. *Academic medicine, 80*(10), S46-S54.

Eva, K. W., & Regehr, G. (2008). “I'll never play professional football” and other fallacies of self-assessment. *Journal of Continuing Education in the Health Professions, 28*(1), 14-19. https://doi.org/10.1002/chp.150

*Eva, K. W., & Regehr, G. (2011). Exploring the divergence between self-assessment and self-monitoring. *Advances in health sciences education, 16*(3), 311-329. https://doi.org/10.1007/s10459-010-9263-2

*Evans, A. W., Leeson, R. M., & Petrie, A. (2007). Reliability of peer and self-assessment scores compared with trainers' scores following third molar surgery. *Medical Education, 41*(9), 866-872. https://doi.org/10.1111/j.1365-2923.2007.02819.x

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of educational research*, *59*(4), 395-430. https://doi.org/10.3102%2F00346543059004395

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of educational research, 70*(3), 287-322. https://doi.org/10.3102%2F00346543070003287

Ferrero, M., Vadillo, M. A., & León, S. P. (2021). A valid evaluation of the theory of multiple intelligences is not yet possible: Problems of methodological quality for

intervention studies. *Intelligence, 88*, 101566. https://doi.org/10.1016/j.intell.2021.101566

*Fieler, G. (2020). Utilization of Video for Competency Evaluation. *Nursing education perspectives, 41*(4), 255-257. https://doi.org/10.1097/01.NEP.0000000000000645

*Fitzgerald, J. T., Gruppen, L. D., & White, C. B. (2000). The influence of task formats on the accuracy of medical students' self-assessments. *Academic Medicine, 75*(7), 737-741.

*Fitzgerald, J. T., White, C. B., & Gruppen, L. D. (2003). A longitudinal study of self-assessment accuracy. *Medical education, 37*(7), 645-649. https://doi.org/10.1046/j.1365-2923.2003.01567.x

*Fitzgerald, K., & Vaughan, B. (2018). Learning through multiple lenses: analysis of self, peer, nearpeer, and faculty assessments of a clinical history-taking task in Australia. *Journal of educational evaluation for health professions*, 15:22. https://doi.org/10.3352/jeehp.2018.15.22

*Frye, A. W., Richards, B. F., Bradley, E. W., & Philp, J. R. (1992). The consistency of students' self-assessments in short-essay subject matter examinations. *Medical Education, 26*(4), 310-316. https://doi.org/10.1111/j.1365-2923.1992.tb00174.x

*Ganni, S., Botden, S. M., Schaap, D. P., Verhoeven, B. H., Goossens, R. H., & Jakimowicz, J. J. (2018). “Reflection-before-practice” improves self-assessment and end-performance in laparoscopic surgical skills training. *Journal of Surgical Education*, *75*(2), 527-533. https://doi.org/10.1016/j.jsurg.2017.07.030

*Geranmayeh, M., Khakbazan, Z., Azizi, F., & Mehran, A. (2020). Effects of Feedback on Midwifery Students’ Self-Assessed Performance and Their Self-Assessment Ability: A Quasi-Experimental Study. *International Quarterly of Community Health Education, 40*(4), 299-305. https://doi.org/10.1177%2F0272684X19885512

*González-Betancor, S. M., Bolívar-Cruz, A., & Verano-Tacoronte, D. (2019). Self-assessment accuracy in higher education: The influence of gender and performance of university students. *Active learning in higher education, 20*(2), 101-114. https://doi.org/10.1177%2F1469787417735604

*Grant, A. L., & Temple-Oberle, C. (2017). Utility of a validated rating scale for self-assessment in microsurgical training. *Journal of Surgical Education, 74*(2), 360-364. https://doi.org/10.1016/j.jsurg.2016.08.017

*Groenendijk, T., Kárpáti, A., & Haanstra, F. (2020). Self-Assessment in Art Education through a Visual Rubric. *International Journal of Art & Design Education, 39*(1), 153-175. https://doi.org/10.1111/jade.12233

*Guest, J., & Riegler, R. (2017). Learning by doing: do economics students self-evaluation skills improve?. *International Review of Economics Education, 24*, 50-64. https://doi.org/10.1016/j.iree.2016.10.002

*Guest, J., & Riegler, R. (2022). Knowing HE standards: how good are students at evaluating academic work?. *Higher Education Research & Development*, *41*(3), 714-728. https://doi.org/10.1080/07294360.2020.1867516

Gutierrez de Blume, A. P. (2022). Calibrating calibration: A meta-analysis of learning strategy instruction interventions to improve metacognitive monitoring accuracy. *Journal of Educational Psychology, 114*(4), 681-700. https://doi.org/10.1037/edu0000674

*Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology, 92*(1), 160. https://doi.apa.org/doi/10.1037/0022-0663.92.1.160

*Hall, S. R., Stephens, J. R., Seaby, E. G., Andrade, M. G., Lowry, A. F., Parton, W. J., Smith, C. F., & Border, S. (2016). Can medical students accurately predict their learning? A study comparing perceived and actual performance in neuroanatomy. *Anatomical Sciences Education, 9*(5), 488-495. https://doi.org/10.1002/ase.1601

*Han, C. (2018). A longitudinal quantitative investigation into the concurrent validity of self and peer assessment applied to English-Chinese bi-directional interpretation in an undergraduate interpreting course. *Studies in Educational Evaluation, 58*, 187-196. https://doi.org/10.1016/j.stueduc.2018.01.001

*Han, C., & Riazi, M. (2018). The accuracy of student self-assessments of English-Chinese bidirectional interpretation: a longitudinal quantitative study. *Assessment & Evaluation in Higher Education, 43*(3), 386-398. https://doi.org/10.1080/02602938.2017.1353062

*Harrington, J. P., Murnaghan, J. J., & Regehr, G. (1997). Applying a relative ranking model to the self-assessment of extended performances. *Advances in Health Sciences Education, 2*(1), 17-25. https://doi.org/10.1023/A:1009782022956

Harris, L. R., & Brown, G. T. L. (2018). *Using self-assessment to improve student learning*. Routledge.

Harzing, A. W. (2007). Publish or Perish. available from https://harzing.com/resources/publish-or-perish.

*Hawkins, S. C., Osborne, A., Schofield, S. J., Pournaras, D. J., & Chester, J. F. (2012). Improving the accuracy of self-assessment of practical clinical skills using video feedback–the importance of including benchmarks. *Medical teacher, 34*(4), 279-284. https://doi.org/10.3109/0142159X.2012.658897

*Herrera-Almario, G. E., Kirk, K., Guerrero, V. T., Jeong, K., Kim, S., & Hamad, G. G. (2016). The effect of video review of resident laparoscopic surgical skills measured by self-and external assessment. *The American Journal of Surgery*, *211*(2), 315-320. https://doi.org/10.1016/j.amjsurg.2015.05.039

*Hewitt, M. P. (2002). Self-evaluation tendencies of junior high instrumentalists. *Journal of research in music education, 50*(3), 215-226. https://doi.org/10.2307%2F3345799

*Hewitt, M. P. (2005). Self-evaluation accuracy among high school and middle school instrumentalists. *Journal of Research in Music Education, 53*(2), 148-161. https://doi.org/10.1177%2F002242940505300205

*Hewitt, M. P. (2011). The impact of self-evaluation instruction on student self-evaluation, music performance, and self-evaluation accuracy. *Journal of Research in Music Education, 59*(1), 6-20. https://doi.org/10.1177%2F0022429410391541

*Hitzeman, C., Gonsalvez, C. J., Britt, E., & Moses, K. (2020). Clinical psychology trainees' self versus supervisor assessments of practitioner competencies. *Clinical Psychologist, 24*(1), 18-29. https://doi.org/10.1111/cp.12183

*Hosein, A., & Harle, J. (2018). The relationship between students' prior mathematical attainment, knowledge and confidence on their self-assessment accuracy. *Studies in Educational Evaluation, 56*, 32-41. https://doi.org/10.1016/j.stueduc.2017.10.008

*Hu, Y., Kim, H., Mahmutovic, A., Choi, J., Le, I., & Rasmussen, S. (2015). Verification of accurate technical insight: a prerequisite for self-directed surgical training. *Advances in Health Sciences Education, 20*(1), 181-191. https://doi.org/10.1007/s10459-014-9519-3

*Hu, Y., Tiemann, D., & Brunt, L. M. (2013). Video self-assessment of basic suturing and knot tying skills by novice trainees. *Journal of surgical education, 70*(2), 279-283. https://doi.org/10.1016/j.jsurg.2012.10.003

*Huth, K. C., Baumann, M., Kollmuss, M., Hickel, R., Fischer, M. R., & Paschos, E. (2017). Assessment of practical tasks in the Phantom course of Conservative Dentistry by pre-defined criteria: a comparison between self-assessment by students and assessment by instructors. *European Journal of Dental Education, 21*(1), 37-45. https://doi.org/10.1111/eje.12176

*Iglesias Pérez, M. C., Vidal-Puga, J., & Pino Juste, M. R. (2022). The role of self and peer assessment in Higher Education. *Studies in Higher Education, 47*(3), 683-692. https://doi.org/10.1080/03075079.2020.1783526

*Iguchi, A., Hasegawa, Y., & Fujii, K. (2020). Student potential for self-assessment in a clinical dentistry practical training course on communication skills. *Medical Science Educator*, *30*, 1503-1513. https://doi.org/10.1007/s40670-020-01061-5

*Jahan, F., Sadaf, S., Bhanji, S., Naeem, N., & Qureshi, R. (2011). Clinical skills assessment: comparison of student and examiner assessment in an objective structured clinical examination. *Education for Health*, *24*(2), 421. Retrieved from https://educationforhealth.net//text.asp?2011/24/2/421/101446

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational research review, 2*(2), 130-144. https://doi.org/10.1016/j.edurev.2007.05.002

*Juhani, T., Mika, H., & Esko, K. (2008). Psychology students' self-assessment of their professional skills. *Nordic Psychology, 60*(4), 267-282. https://doi.org/10.1027/1901-2276.60.4.267

*Kachingwe, A. F., Phillips, B., & Beling, J. (2015). Videotaping practical examinations in physical therapist education: does it foster student performance, self-assessment, professionalism, and improve instructor grading?. *Journal of Physical Therapy Education*, *29*(1), 25-33. Retrieved from https://journals.lww.com/jopte/Fulltext/2015/29010/Videotaping_Practical_Examinations_in_Physical.6.aspx

*Karnilowicz, W. (2012). A comparison of self-assessment and tutor assessment of undergraduate psychology students. *Social Behavior and Personality: an international journal, 40*(4), 591-604. Retrieved from

https://vuir.vu.edu.au/23230/2/A_COMPARISON_OF_SELF-ASSESSMEN.pdf

*Keane, L., & Griffin, C. P. (2018). Assessing self-assessment: can age and prior literacy attainment predict the accuracy of children's self-assessments in literacy?. *Irish Educational Studies, 37*(1), 127-147. https://doi.org/10.1080/03323315.2018.1449001

*Kearney, S., Perkins, T., & Kennedy-Clark, S. (2016). Using self-and peer-assessments for summative purposes: analysing the relative validity of the AASL (Authentic Assessment for Sustainable Learning) model. *Assessment & Evaluation in Higher Education, 41*(6), 840-853. https://doi.org/10.1080/02602938.2015.1039484

*Kilic, D. (2016). An Examination of Using Self-, Peer-, and Teacher-Assessment in Higher Education: A Case Study in Teacher Education. *Higher Education Studies*, *6*(1), 136-144. https://doi.org/10.5539/hes.v6n1p136

Kitsantas, A., & Zimmerman, B. J. (2006). Enhancing self-regulation of practice: The influence of graphing and self-evaluative standards. *Metacognition and Learning,* 201-212. https://doi.org/10.1007/s11409-006-9000-7

Klenowski, V. (1995). Student self-evaluation processes in student-centred teaching and learning contexts of Australia and England. *Assessment in Education: Principles, Policy & Practice, 2*(2), 145-163. https://doi.org/10.1080/0969594950020203

*Kostka, M. J. (1997). Effects of self-assessment and successive approximations on "knowing" and "valuing" selected keyboard skills. *Journal of Research in Music Education, 45*(2), 273-281. https://doi.org/10.2307%2F3345586

*Kostons, D., van Gog, T., & Paas, F. (2010). Self-assessment and task selection in learner-controlled instruction: Differences between effective and ineffective learners. *Computers & Education, 54*(4), 932-940. https://doi.org/10.1016/j.compedu.2009.09.025

Krebs, R., Rothstein, B., & Roelle, J. (2022, 2022/05/19). Rubrics enhance accuracy and reduce cognitive load in self-assessment. *Metacognition and Learning*. https://doi.org/10.1007/s11409-022-09302-1

*Kun, A. I. (2016). A comparison of self versus tutor assessment among Hungarian undergraduate business students. *Assessment & Evaluation in Higher Education, 41*(3), 350-367. https://doi.org/10.1080/02602938.2015.1011602

*Kustritz, M. V., Molgaard, L. K., & Rendahl, A. (2011). Comparison of student self-assessment with faculty assessment of clinical competence. *Journal of veterinary medical education, 38*(2), 163-170. https://doi.org/10.3138/jvme.38.2.163

*Kwasnik, E. M., & Carter, J. (1999). To see ourselves as others see us: self-assessment in surgical residency. *Current Surgery*, *56*(3), 145-148. https://doi.org/10.1016/S0149-7944(99)00041-0

*Langan, A. M., Shuker, D. M., Cullen, W. R., Penney, D., Preziosi, R. F., & Wheater, C. P. (2008). Relationships between student characteristics and self-, peer and tutor evaluations of oral presentations. *Assessment & Evaluation in Higher Education, 33*(2), 179-190. https://doi.org/10.1080/02602930701292498

*Lappin-Fortin, K., & Rye, B. J. (2014). The use of pre-/posttest and self-assessment tools in a French pronunciation course. *Foreign Language Annals, 47*(2), 300-320. https://doi.org/10.1111/flan.12083

*Lau, E., Dolovich, L., & Austin, Z. (2007). Comparison of self, physician, and simulated patient ratings of pharmacist performance in a family practice simulator. *Journal of interprofessional care, 21*(2), 129-140. https://doi.org/10.1080/13561820601133981

*Lavrysh, Y. (2016). Peer and self-assessment at ESP classes: case study. *Advanced education,* (6), 60-68. https://doi.org/10.20535/2410-8286.85351

*Leach, L. (2012). Optional self-assessment: some tensions and dilemmas. *Assessment & Evaluation in Higher Education, 37*(2), 137-147. https://doi.org/10.1080/02602938.2010.515013

León, S. P., & García-Martínez, I. (2021). Feedback and evaluative experience as decisive factors in student self-regulation. *Publicaciones, 51*(1), 303–316. https://doi.org/10.30827/publicaciones.v51i1.20738

*León, S. P., Pantoja Vallejo, A., & Nelson, J. B. (2021). Variability in the accuracy of self-assessments among low, moderate, and high performing students in university education. *Practical Assessment, Research, and Evaluation*, *26*(1), 16. https://doi.org/10.7275/6q91-az58

*Lew, M. D., Alwis, W. A. M., & Schmidt, H. G. (2010). Accuracy of students' self-assessment and their beliefs about its utility. *Assessment & Evaluation in Higher Education, 35*(2), 135-156. https://doi.org/10.1080/02602930802687737

*Lind, D. S., Rekkas, S., Bui, V., Lam, T., Beierle, E., & Copeland Iii, E. M. (2002). Competency-based student self-assessment on a surgery rotation. *Journal of Surgical Research*, *105*(1), 31-34. https://doi.org/10.1006/jsre.2002.6442

*Lindblom-ylänne, S., Pihlajamäki, H., & Kotkas, T. (2006). Self-, peer-and teacher-assessment of student essays. *Active learning in higher education, 7*(1), 51-62. https://doi.org/10.1177%2F1469787406061148 *Liu, C. S., Wang, Y. M., & Lin, H. N. (2021). An 8-year retrospective survey of assessment in postgraduate dental training in complicated tooth extraction competency. *Journal of Dental Sciences*, *16*(3), 891-898. https://doi.org/10.1016/j.jds.2020.11.006

*López-Pastor, V. M., Fernández-Balboa, J. M., Santos Pastor, M. L., & Fraile Aranda, A. (2012). Students' self-grading, professor's grading and negotiated final grading at three university programmes: analysis of reliability and grade difference ranges and tendencies. *Assessment & Evaluation in Higher Education, 37*(4), 453-464. https://doi.org/10.1080/02602938.2010.545868

*Lundquist, L. M., Shogbon, A. O., Momary, K. M., & Rogers, H. K. (2013). A comparison of students' self-assessments with faculty evaluations of their communication skills. *American Journal of Pharmaceutical Education, 77*(4). https://doi.org/10.5688/ajpe77472

*Machado, J. L. M., Machado, V. M. P., Grec, W., Bollela, V. R., & Vieira, J. E. (2008). Self-and peer assessment may not be an accurate measure of PBL tutorial process. *BMC medical education, 8*(1), 1-6. https://doi.org/10.1186/1472-6920-8-55

*Marks, M. B., Haug, J. C., & Hu, H. (2018). Investigating undergraduate business internships: Do supervisor and self-evaluations differ?. *Journal of Education for Business, 93*(2), 33-45. https://doi.org/10.1080/08832323.2017.1414025

*Martínez, V., Mon, M. A., Álvarez, M., Fueyo, E., & Dobarro, A. (2020). E-self-assessment as a strategy to improve the learning process at university. *Education Research International, 2020*. https://doi.org/10.1155/2020/3454783

*Máté, D., Bács, Z., & Takács, V. L. (2017). Analyzing the Implementation of an ERP System by Self-Assessment in Higher Education. *Acta Didactica Napocensia*, *10*(2), 45-56. Retrieved from https://files.eric.ed.gov/fulltext/EJ1156615.pdf

*Máté, D., & Darabos, É. (2017). Measuring the accuracy of self-assessment among undergraduate students in higher education to enhance competitiveness. *Journal of Competitiveness*, *9*(2). https://doi.org/10.7441/joc.2017.02.06

*Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language testing, 26*(1), 075-100. https://doi.org/10.1177%2F0265532208097337

*McKevitt, C. T. (2016). Engaging students with self-assessment and tutor feedback to improve performance and support assessment capacity. *Journal of University Teaching & Learning Practice, 13*(1), 2. https://doi.org/10.53761/1.13.1.2

*Miller, T. M., & Geraci, L. (2011). Training metacognition in the classroom: The influence of incentives and feedback on exam predictions. *Metacognition and Learning, 6*(3), 303-314. https://doi.org/10.1007/s11409-011-9083-7

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Prisma Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS medicine, 6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

*Munoz, A., & Álvarez, M. E. (2007). Students' objectivity and perception of self assessment in an EFL classroom. *The Journal of Asia TEFL, 4*(2), 1-25. Retrieved from https://www.academia.edu/download/38190204/4-2-01Asia_TEFL_2_.pdf

*Nederhand, M. L., Tabbers, H. K., Jongerling, J., & Rikers, R. M. (2020). Reflection on exam grades to improve calibration of secondary school students: a longitudinal study. *Metacognition and Learning, 15*(3), 291-317. https://doi.org/10.1007/s11409-020-09233-9

*Nejad, A. M., & Mahfoodh, O. H. A. (2019). Assessment of Oral Presentations: Effectiveness of Self-, Peer-, and Teacher Assessments. *International Journal of Instruction, 12*(3), 615-632. https://doi.org/10.29333/iji.2019.12337a

*Nisly, S. A., Sebaaly, J., Fillius, A. G., Haltom, W. R., & Dinkins, M. M. (2020). Changes in pharmacy students' metacognition through self-evaluation during advanced pharmacy practice experiences. *American journal of pharmaceutical education, 84*(1). https://doi.org/10.5688/ajpe7489

*Oh, S. L., Liberman, L., & Mishler, O. (2018). Faculty calibration and students' self-assessments using an instructional rubric in preparation for a practical

examination. *European Journal of Dental Education, 22*(3), e400-e407. https://doi.org/10.1111/eje.12318

*Öhrstedt, M., & Lindfors, P. (2019). First-semester students' capacity to predict academic achievement as related to approaches to learning. *Journal of Further and Higher Education, 43*(10), 1420-1432. https://doi.org/10.1080/0309877X.2018.1490950

*Osborne, A. J., Hawkins, S. C., Pournaras, D. J., Chandratilake, M., & Welbourn, R. (2014). An evaluation of operative self-assessment by UK postgraduate trainees. *Medical Teacher, 36*(1), 32-37. https://doi.org/10.3109/0142159X.2013.836268

*Osterhage, J. L., Usher, E. L., Douin, T. A., & Bailey, W. M. (2019). Opportunities for self-evaluation increase student calibration in an introductory biology course. *CBE—Life Sciences Education, 18*(2), ar16. https://doi.org/10.1187/cbe.18-10-0202

Panadero, E. (2023). Toward a paradigm shift in feedback research: Five further steps influenced by Self-Regulated learning theory. *Educational Psychologist.*

Panadero, E., & Alonso-Tapia, J. (2013). Self-assessment: Theoretical and practical connotations. When it happens, how is it acquired and what to do to develop it in our students. *Electronic Journal of Research in Educational Psychology, 11*(2), 551-576. https://doi.org/http://dx.doi.org/10.14204/ejrep.30.12200

Panadero, E., & Alqassab, M. (2019). An empirical review of anonymity effects in peer assessment, peer feedback, peer review, peer evaluation and peer grading. *Assessment & Evaluation in Higher Education*, *44:8*, 1253-1278. https://doi.org/10.1080/02602938.2019.1600186

Panadero, E., Brown, G. T., & Strijbos, J. W. (2016). The future of student self-assessment: A review of known unknowns and potential directions. *Educational psychology review, 28*(4), 803-830. https://doi.org/10.1007/s10648-015-9350-2

Panadero, E., Fernández-Ruiz, J., & Sánchez-Iglesias, I. (2020). Secondary education students' self-assessment: The effects of feedback, subject matter, year level, and gender. *Assessment in Education: Principles, Policy & Practice, 27*(6), 607-634. https://doi.org/10.1080/0969594X.2020.1835823

Panadero, E., García-Pérez, D., Fernández-Ruiz, J., Fraile, J., Sánchez-Iglesias, I., & Brown, G. T. L. (2022). University students' strategies and criteria during self-

assessment: Instructor's feedback, rubrics, and year level effects. *European Journal of Psychology of Education*. https://doi.org/10.1007/s10212-022-00639-4

Panadero, E., Jonsson, A., & Botella, J. (2017). Effects of self-assessment on self-regulated learning and self-efficacy: Four meta-analyses. *Educational Research Review, 22,* 74-98. https://doi.org/10.1016/j.edurev.2017.08.004

*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. https://doi.org/10.1080/0969594X.2013.877872

Panadero, E., Romero, M., & Strijbos, J. W. (2013). The impact of a rubric and friendship on construct validity of peer assessment, perceived fairness and comfort, and performance. *Studies in Educational Evaluation, 39*(4), 195–203. https://doi.org/10.1016/j.stueduc.2013.10.005.

*Papinczak, T., Young, L., Groves, M., & Haynes, M. (2007). An analysis of peer, self, and tutor assessment in problem-based learning tutorials. *Medical teacher, 29*(5), e122-e132. https://doi.org/10.1080/01421590701294323

*Partido, B. B., & Henderson, R. (2021). Impact of ergonomic training on posture utilizing photography and self-assessments among dental hygiene students and practitioners. *American Dental Hygienists' Association*, *95*(3), 33-41. Retrieved from https://jdh.adha.org/content/95/3/33.full

*Partido, B. B. (2020). Longitudinal effects of utilising photography on the accuracy of ergonomic self-assessments amongst dental hygiene students. E*uropean Journal of Dental Education, 24*(1), 63-70. https://doi.org/10.1111/eje.12468

*Pawluk, S. A., Zolezzi, M., & Rainkie, D. (2018). Comparing student self-assessments of global communication with trained faculty and standardized patient assessments. *Currents in Pharmacy Teaching and Learning, 10*(6), 779-784. https://doi.org/10.1016/j.cptl.2018.03.012

*Peyre, S. E., MacDonald, H., Al-Marayati, L., Templeman, C., & Muderspach, L. I. (2010). Resident self-assessment versus faculty assessment of laparoscopic technical skills using a global rating scale. *International Journal of Medical Education, 1*, 37-41. https://doi.org/10.5116/ijme.4bf1.c3c1

*Pilotti, M. A., El Alaoui, K., Mulhem, H., & Al Kuhayli, H. A. (2019). The illusion of knowing in college: A field study of students with a teacher-centered educational past. *Europe's Journal of Psychology, 15*(4), 789. https://doi.org/10.5964%2Fejop.v15i4.1921

*Piper, K., Morphet, J., & Bonnamy, J. (2019). Improving student-centered feedback through self-assessment. *Nurse Education Today, 83*, 104193. https://doi.org/10.1016/j.nedt.2019.08.011

*Pui, P., Yuen, B., & Goh, H. (2021). Using a criterion-referenced rubric to enhance student learning: a case study in a critical thinking and writing module. *Higher Education Research & Development, 40*(5), 1056-1069. https://doi.org/10.1080/07294360.2020.1795811

Raaijmakers, S. F., Baars, M., Paas, F., van Merriënboer, J. J., & van Gog, T. (2019). Effects of self-assessment feedback on self-assessment and task-selection accuracy. Metacognition and Learning, 14(1), 21–42. https://doi.org/10.1007/s11409-019-09189-5

*Reitmeier, C. A., & Vrchota, D. A. (2009). Self-assessment of oral communication presentations in food science and nutrition. *Journal of Food Science Education, 8(*4), 88-92. https://doi.org/10.1111/j.1541-4329.2009.00080.x

*Reuland, D. S., Frasier, P. Y., Olson, M. D., Slatt, L. M., Aleman, M. A., & Fernandez, A. (2009). Accuracy of self-assessed Spanish fluency in medical students. *Teaching and Learning in Medicine, 21*(4), 305-309. https://doi.org/10.1080/10401330903228489

*Rezler, A. G. (1989). Self-assessment in problem-based groups. *Medical Teacher, 11*(2), 151-156. https://doi.org/10.3109/01421598909146318

*Ricciotti, H. A., Dodge, L. E., Head, J., Atkins, K. M., & Hacker, M. R. (2012). A novel resident-as-teacher training program to improve and evaluate obstetrics and gynecology resident teaching skills. *Medical Teacher*, *34*(1), e52-e57. https://doi.org/10.3109/0142159X.2012.638012

*Ritchie, S. M. (2016). Self-assessment of video-recorded presentations: Does it improve skills?. *Active Learning in Higher Education, 17*(3), 207-221. https://doi.org/10.1177%2F1469787416654807

*Rudy, D. W., Fejfar, M. C., Griffith III, C. H., & Wilson, J. F. (2001). Self-and peer assessment in a first-year communication and interviewing course. *Evaluation &*

the Health Professions, *24*(4), 436-445. https://journals.sagepub.com/doi/pdf/10.1177/016327870102400405

*Sadosty, A. T., Bellolio, M. F., Laack, T. A., Luke, A., Weaver, A., & Goyal, D. G. (2011). Simulation-based emergency medicine resident self-assessment. *The Journal of emergency medicine*, *41*(6), 679-685. https://doi.org/10.1016/j.jemermed.2011.05.041

*Salehi, M., & Masoule, Z. S. (2017). An investigation of the reliability and validity of peer, self-, and teacher assessment. *Southern African Linguistics and Applied Language Studies*, *35*(1), 1-15. https://doi.org/10.2989/16073614.2016.1267577

*San Diego, J. P., Newton, T., Quinn, B. F. A., Cox, M. J., & Woolford, M. J. (2014). Levels of agreement between student and staff assessments of clinical skills in performing cavity preparation in artificial teeth. *European Journal of Dental Education, 18*(1), 58-64. https://doi.org/10.1111/eje.12059

*Sanderson, T. R., Kearney, R. C., Kissell, D., & Salisbury, J. (2016). Evaluating student self-assessment through video-recorded patient simulations. *American Dental Hygienists' Association*, *90*(4), 257-262. Retrieved from https://jdh.adha.org/content/90/4/257

Sanchez, C. E., Atkinson, K. M., Koenka, A. C., Moshontz, H., & Cooper, H. (2017). Self-grading and peer-grading for formative and summative assessments in 3rd through 12th grade classrooms: A meta-analysis. *Journal of Educational Psychology, 109*(8), 1049. https://doi.apa.org/doi/10.1037/edu0000190

*Sasmaz Oren, F. (2012). The effects of gender and previous experience on the approach of self and peer assessment: A case from Turkey. *Innovations in Education and Teaching International, 49*(2), 123-133. https://doi.org/10.1080/14703297.2012.677598

*Sendlhofer, G., Pregartner, G., Gombotz, V., Leitgeb, K., Tiefenbacher, P., Jantscher, L., ... & Brunner, G. (2019). A new approach of assessing patient safety aspects in routine practice using the example of “doctors handwritten prescriptions”. *Journal of clinical nursing, 28*(7-8), 1242-1250. https://doi.org/10.1111/jocn.14736

Silver, N. C., & Dunlap, W. P. (1987). Averaging correlation coefficients: Should Fisher’s z transformation be used? *Journal of Applied Psychology, 72*, 146 –148. http://dx.doi.org/10.1037/0021-9010.72.1.146

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *P*-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143*, 534 –547. http://dx.doi.org/10.1037/a0033242

Sitzmann, T., Ely, K., Brown, K. G., & Bauer, K. N. (2010). Self-assessment of knowledge: a cognitive learning or affective measure? *Academy of Management Learning & Education, 9*(2), 169–191. https://doi.org/10.5465/amle.9.2.zqr169

*Spoto-Cannons, A. C., Isom, D. M., Feldman, M., Zwygart, K. K., Mhaskar, R., & Greenberg, M. R. (2019). Differences in medical student self-evaluations of clinical and professional skills. *Advances in medical education and practice, 10*, 835. https://doi.org/10.2147%2FAMEP.S222774

*Srikumaran, D., Tian, J., Ramulu, P., Boland, M. V., Woreta, F., Wang, K. M., & Mahoney, N. (2019). Ability of ophthalmology residents to self-assess their performance through established milestones. *Journal of surgical education, 76(*4), 1076-1087. https://doi.org/10.1016/j.jsurg.2018.12.004

*Stahl, C. C., Jung, S. A., Rosser, A. A., Kraut, A. S., Schnapp, B. H., Westergaard, M., Hamedani, A. G., Minter, R. M., & Greenberg, J. A. (2020). Entrustable professional activities in general surgery: trends in resident self-assessment. *Journal of surgical education, 77*(6), 1562-1567. https://doi.org/10.1016/j.jsurg.2018.12.004

*Stauffer, L. K. (2011). ASL students' ability to self assess ASL competency. *Journal of Interpretation*, *21*(1), 7. Retrieved from https://digitalcommons.unf.edu/joi/vol21/iss1/7

Stone, N. J. (2000). Exploring the relationship between calibration and self-regulated learning. *Educational Psychology Review, 12*(4), 437-475. https://doi.org/10.1023/a:1009084430926

*Storjohann, T., Pogge, E., Peckham, A., Raney, E., & Barletta, J. F. (2019). Evaluation of a peer-and self-grading process for clinical writing assignments. *Currents in Pharmacy Teaching and Learning, 11*(10), 979-986. https://doi.org/10.1016/j.cptl.2019.06.003

*Swank, J. M. (2014). Assessing counseling competencies: A comparison of supervisors' ratings and student supervisees' self-ratings. *Counseling Outcome Research and Evaluation, 5*(1), 17-27. https://doi.org/10.1177%2F2150137814529147

Taras, M. (2010). Student self-assessment: processes and consequences. *Teaching in Higher Education, 15*(2), 199–209. https://doi.org/10.1080/13562511003620027

Tejeiro, R. A., Gomez-Vallecillo, J. L., Romero, A. F., Pelegrina, M., Wallace, A., & Emberley, E. (2012). Summative self-assessment in higher education: Implications of its counting towards the final mark. *Electronic Journal of Research in Educational Psychology, 10*(2), 789-812. Retrieved from http://investigacion-psicopedagogica.org/revista/articulos/27/english/Art_27_707.pdf

*Tousignant, M., & DesMarchais, J. E. (2002). Accuracy of student self-assessment ability compared to their own performance in a problem-based learning medical program: a correlation study. *Advances in Health Sciences Education, 7*(1), 19-27. https://doi.org/10.1023/A:1014516206120

*Trofimovich, P., Isaacs, T., Kennedy, S., Saito, K., & Crowther, D. (2016). Flawed self-assessment: Investigating self-and other-perception of second language speech. *Bilingualism: Language and Cognition, 19*(1), 122-140. https://doi.org/10.1017/S1366728914000832

*Tuncer, D., Arhun, N., Yamanel, K., Çelik, Ç., & Dayangaç, B. (2015). Dental students' ability to assess their performance in a preclinical restorative course: comparison of students' and faculty members' assessments. *Journal of dental education*, *79*(6), 658-664. https://doi.org/10.1002/j.0022-0337.2015.79.6.tb05938.x

*Ünaldı, İ. (2016). Self and teacher assessment as predictors of proficiency levels of Turkish EFL learners. *Assessment & evaluation in higher education, 41*(1), 67-80. https://doi.org/10.1080/02602938.2014.980223

*van Hattum-Janssen, N., & Lourenço, J. M. (2008). Peer and self-assessment for first-year students as a tool to improve learning. *Journal of professional issues in engineering education and practice, 134*(4), 346-352. https://doi.org/10.1061/(ASCE)1052-3928(2008)134:4(346)

Viechtbauer, W. (2010). Conducting meta–analyses in R with the metafor package. *Journal of Statistical Software, 36,* 1–48. https://doi.org/10.18637/jss.v036.i03

Viechtbauer, W., & Cheung, M. W.-L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods, 1*(2), 112–125. https://doi.org/10.1002/jrsm.11

*Wagner, M. L., Suh, D. C., & Cruz, S. (2011). Peer-and self-grading compared to faculty grading. *American journal of pharmaceutical education, 75*(7). https://doi.org/10.5688/ajpe757130

Ward, M., Gruppen, L., & Regehr, G. (2002). Measuring self-assessment: current state of the art. *Advances in Health Sciences Education, 7*(1), 63–81. https://doi.org/10.1023/A:1014585522084

*Ward, M., MacRae, H., Schlachta, C., Mamazza, J., Poulin, E., Reznick, R., & Regehr, G. (2003). Resident self-assessment of operative performance. *The American journal of surgery*, *185*(6), 521-524. https://doi.org/10.1016/S0002-9610(03)00069-2

*Weiss, P. M., Koller, C. A., Hess, L. W., & Wasser, T. (2005). How do medical student self-assessments compare with their final clerkship grades?. *Medical teacher, 27*(5), 445-449. https://doi.org/10.1080/01421590500046999

*Wettergreen, S. A., Brunner, J., Linnebur, S. A., Borgelt, L. M., & Saseen, J. J. (2018). Comparison of faculty assessment and students' self-assessment of performance during clinical case discussions in a pharmacotherapy capstone course. *Medical teacher, 40*(2), 193-198. https://doi.org/10.1080/0142159X.2017.1397271

Wiliam, D. (2011). What is assessment for learning?. *Studies in educational evaluation, 37*(1), 3-14. https://doi.org/10.1016/j.stueduc.2011.03.001

*Wilson, J., & Wright, C. R. (1993). The Predictive Validity of Student Self-Evaluations, Teachers' Assessments, and Grades for Performance on the Verbal Reasoning and Numerical Ability Scales of the Differential Aptitude Test for a Sample of Secondary School Students Attj7Ending Rural Appalachia Schools. *Educational and psychological measurement, 53*(1), 259-270. https://doi.org/10.1177%2F0013164493053001029

*Wong, H. M. (2016). I can assess myself: Singaporean primary students' and teachers' perceptions of students' self-assessment ability. *Education 3-13, 44*(4), 442-457. https://doi.org/10.1080/03004279.2014.982672

Yan, Z. (2018). Student self-assessment practices: The role of gender, school level and goal orientation. *Assessment in Education: Principles, Policy & Practice, 25*(2), 183-199. https://doi.org/10.1080/0969594X.2016.1218324

Table 1. Articles that met inclusion and quality criteria.

Author (year)	Aim of the study	Sample size	Age (mean)	Area of knowledge	Educational level	Procedure	Self-assessment task	Measurement of task performance	Analysis	Results
Abadel et al. (2013)	Assess SA clinical competency.	105	(28.8 years)	Medicine	Higher education	Medical students' scores on a questionnaire about their medical competences were compared with the scores given by teachers. Sociometric variables were assessed.	Clinical competency	Assessment questionnaire	Paired t-test	Large discrepancy between SA and experts' assessment of graduates' clinical competence. Graduates tend to overestimate their competence.
Abate et al. (2007)	To describe the use of student SA as a measure of the effectiveness of a drug information advanced pharmacy practice experience (APPE).	56	n.s.	Pharmacy	Higher education	An SA questionnaire is administered before and after the APPE course. SA scores are compared for gender, and for high and low levels in the results of a course on medical literature evaluation (Medical Literature Evaluation).	Drug information	Drug Information Skills Self-Assessment Survey instrument	Wilcoxon signed-rank	There are differences Pre-Post SA. There are differences in Pre-Post SA for gender and for Level of competencies in assessment. There are no differences between Post SA and Standard.
Abdalla et al. (2021)	Assess the correlation between SA, visual and digital grades and the reliability of the software.	56	22-39 years	Odontology	Higher education	He training lasted four three-hour sessions. All students were fully trained in the use of the assessment software and received feedback from two sources: faculty members and the assessment software. They were instructed in SA.	Visual and digital grades	Tooth # 23	T tests and Correlations	Correlations between SA and visual and digital grades improved with 19 versus 23. A near-perfect correlation was found between grades at the first and second digital grading sessions.
Abeyaratne et al. (2022)	The objectives for this study were to determine how well third-year pharmacy students self-assess	162	n.s.	Pharmacy	Higher education	A quasi-experimental one-group pre-/post-test design was conducted with third-year pharmacy students.	Therapeutics course	Standard model of patient care to identify medication-related problems (MRPs)	T test and Correlations	On average, students demonstrated poor self-evaluation skills and underestimated themselves by 4.9%
Admiraal et al. (2015)	Reliability of both the SA the peer assessments performance in Massive Open Online Course (MOOC).	410	n.s.	Medicine	Higher education	For 3 MOOCs (duration between 5 and 8 weeks each). There were learners who registered and learners who did not register their progress (those who registered showed better scores in the final exam).	Content and structure of the essay	Weekly quiz (multiple-choice)	Stepwise regression analyses	SA does not explain test differences. Does not find that the three types of assessments have a beneficial effect on final exams.
Agrawal et al. (2012)	Evaluate aspects of a testing experience most influence self-regulated learning behaviour among medical students.	67	(25 years)	Medicine	Higher education	Students answer 10 clinical confidence questions, after review they are asked to estimate how good they are. This process is repeated several times.	Six clinical domains	Computer-based, multiple-choice test (10 multiple-choice questions)	Correlations	Weak SA accuracy at 1st attempt 0.28, increases after feedback at 2nd attempt 0.39.
Aiko (2018)	Evaluate the accuracy of kanji students analysed their kanji skills and analyse possible factors affecting SA accuracy.	31	n.s.	Languages	Primary education	Assess how children evaluate their competence in different areas of Japanese (reading, writing, etc...)	Kanji Reading and Writing	kanji tests and questionnaires	Correlations	Greater precision for the more competent and overestimation for the less competent.
Aitken et al. (2018)	Create an essay model for students to rehearse their SA skills.	71	n.s.	Design	Higher education	Mock test using a computer programme (REVIEW) that assesses writing ability, criteria and feedback are provided.	Academic writing	REVIEW software	n.s.	The most accurate students are the ones with the highest scores.

Table 1 (*continued*)

Author (year)	Aim of the study	Sample size	Age (mean)	Area of knowledge	Educational level	Procedure	Self-assessment task	Measurement of task performance	Analysis	Results
Akkus et al. (2017)	Evaluates the impact of microteaching on the skills of future chemistry teachers.	31	n.s.	Chemistry	Higher education	Pre-post design without controls. They do a presentation + recording of how they did it + SA. This is repeated a second time.	Communication skills performance, illustrated talk performance, and process skill lesson performance.	Questionnaire consists of three rating scales.	Paired t-test	Differences between PRE and POST are analysed for all evaluation measures. All improve significantly.
Akyuz (2018)	Analysing lesson plans collected from preservice mathematics teachers obtained from a technology-integration course over a period of five years by using a novel instrument.	138	n.s.	Mathematics	Higher education	During a course, a questionnaire based on the variables proposed by the Technological, Pedagogical, And Content Knowledge (TPACK) model is used to measure both course performance and SA.	Technical and pedagogical competences of the TPACK model for geometry teaching	TPACK self-assessment surveys	Correlations	It analyses the validity of the tool, with a good fit and validity. The difference between SA and STD is not analysed.
Alameddine et al. (2018)	Assess the feasibility of a video-based coaching intervention to improve the suturing skills of fourth-year medical students.	16	n.s.	Medicine	Higher education	Experimental vs Control. Students in the intervention group received a structured coaching session between consecutive suturing tasks, whereas students in the control group did not. Following each suturing task, students were asked to self-assess their performance and provide feedback.	Suturing skills (bimanual dexterity, efficiency, tissue handling, level of difficulty, and consistency)	Global Operative Assessment of Laparoscopic Skills	n.s.	They analyse the results between experimental and control for the medical skill (surgery), finding hardly any differences between groups. For the Experimental in SA they showed subjective improvement. None of the comparisons were significant
Albanese et al. (2006)	Assess if students inaccurately assess their own skills, especially high- or low-performers on exams.	113	n.s.	Medicine	Higher education	Group students in a class into 5 groups according to their results in 4 different grades (performance). They evaluate the accuracy of SA according to the group (performance).	Basic principles of infection and the body's defences	The final exam consisted of 100 multiple choice questions	Kruskal–Wallace ANOVA	Results found that students accurately assessed their percent correct, but inaccurately assessed their percentile rank. No statistically significant differences existed between the true and false-low subgroups nor the true- and false-high subgroups.
Alfakhry et al. (2022)	To assess the impact of deliberate training on the concordance between student self-assessment and teacher evaluation	16	22-23	Dental	Postgraduate	Students and instructors are instructed in SSA and the instruments. Students had to completed the SSA form after the procedure, while instructors assess students during the procedure.	Sugical skills	Direct Observation of Procedural Skills (DOPS),	Paired t-testing	Bias in SSA decreased consistently as time goes on.
Ammentorp et al. (2013)	Compare the medical students perceived self-efficacy and the evaluation of the observers and patients.	73	n.s.	Medicine	Higher education	The students were presented with paediatric and obstetric cases at 10 written and oral stations. SA, observer and patient assessment were compared.	Specific communication skills.	Calgary-Cambridge Observation Guide Checklist (12 items on a Likert scale from 1-5)	Rasch model	This study showed that students scored their communication skills lower compared to observers or simulated patients.
Andoh et al. (2008)	To find out whether students are able to evaluate their own work as their tutors would do, and to find out whether there are gender differences.	56	n.s.	Law	Higher education	During a seminar, they had to do a task and self-assess their performance grade according to a script. Their scores are compared with those of the teacher.	Contents law	Examen	n.s.	High level of agreement between SA and teacher. Positive relationship between SA and performance. Better SA performance of females.

Table 1 (*continued*)

Author (year)	Aim of the study	Sample size	Age (mean)	Area of knowledge	Educational level	Procedure	Self-assessment task	Measurement of task performance	Analysis	Results
Aryadoust (2015)	Evaluate accuracy of self- and peer assessments of oral presentations.	40	n.s.	Sciences	Higher education	After a 12-week course on oral presentation, SA and PEER were requested and compared with STD.	English oral presentation skills.	Tertiary-level English oral presentation scale measured by 18 items.	Correlations	Self-, peer, and tutor assessments had low to medium correlations on the subscales, and a significant difference was found between the assessments.
Ashton (2014)	Compare secondary school learner SA of reading proficiency in German, Japanese and Urdu for the multilingual assessment scheme.	439	(13.6 years)	Languages	Secondary education	A language test was administered, with 4 different levels of difficulty.	Language skills	Learner self-assessment survey + Asset Languages test data	Correlations	Good correlation between SA and professor scores.
Austin et al. (2007)	To evaluate the accuracy of SA skills of senior-level bachelor of science pharmacy students.	80	n.s.	Sciences	Higher education	Comparisons of pharmacy SA with weighted average assessments of peers, standardized patients, and pharmacy instructors was used.	Verbal/nonverbal communication; degree of focus; logic; empathy; and overall performance.	Questionnaire, 5-point Likert scale.	T-test	Differences between SA and external assessments were found across all performance quartiles. The bad students were the most inaccurate and as the students got better the pressure increased until SA and STD coincided.
Baars et al. (2014)	Assess whether SA accuracy can be improved by training (Experiment 1 and 2) or by providing standard measurements (Experiment 2).	177	(13.56 years)	Biology	Secondary education	2 sessions, 70 minutes: SA training (Experiment 1); 5 sessions of 70 minutes: SA training + indicator guides (Experiment 2).	Troubleshooting and Judgments of Learning	Assessment scales and Exam	T-test	SA accuracy does not improve with training (Experiment 1 and 2). Providing standard measures improves SA
Baecher et al. (2013)	The influence of video models on teacher candidates' capacity to self-evaluate their teaching performance in early fieldwork.	31	24-45	Languages	Higher education	Two groups, Pre-post. After providing small group tutorials (35 hours), they pre-submitted a mini-lesson. One received a video to model and the other received only criteria. Both used a rubric.	EFE observation	Rubric	Correlations	Introduction of video models reduced inflation of scores in self-evaluation and enhanced candidates' understanding of the expectations for the performance assessment of teaching.
Balch (1992)	Provide practical information to teachers regarding differences in students' self-assessment styles.	90	n.s.	Psychology	Higher education	Students are asked to predict their score before taking a test, and are asked again after taking the test. Three groups are made depending on the exam score.	Psychology contents	Multiple-choice final exam	T-test	Below average groups overestimated in pre and post, average groups overestimated only in pre but not in post, and above average groups were accurate in pre and post.
Baleghizadeh et al. (2014)	How Iranian EFL learners developed the ability to self-assess their writings through having access to the rater's scores.	15	(20 years)	Languages	Foreign Languages education	Participants were supervised for four weeks as they underwent their first self-assessment experience. They were provided with a detailed evaluation sheet, and they were able to access the scores assigned by the teacher.	English as a Foreign Language Writing	Detailed evaluation sheet of 5 components	Correlations and T-test	The results indicated a high correlation between self-assessment and teacher evaluation. Over the 4 cycles the correlation between SA and STD increased.
Ballantine et al. (2007)	Self-assessment accuracy as a measure of computer competence.	123	n.s.	Business	Higher education	To evaluate the reliability of self-assessed computer competence, the scores achieved by students in self-assessed computer competence tests are compared with scores achieved in objective tests	Computer competence	Questionnaire, five-point Likert scale	Wilcoxon matched-pairs signed-ranks test	The results reveal a statistically significantly over-estimation of computer competence among the students surveyed.

Table 1 (*continued*)

Author (year)	Aim of the study	Sample size	Age (mean)	Area of knowledge	Educational level	Procedure	Self-assessment task	Measurement of task performance	Analysis	Results
Bergee (1993)	Do SA of applied brass jury performances correlate with faculty and peer evaluations?	5	n.s.	Music	Music education	Different instructors were asked to evaluate with a scale different performance, SA, PEER and STD were asked at three different times.	Music skills	Author-constructed Brass Performance Rating Scale (BPRS).	Correlations	Low correlation at two of the three points in time between STD and SA
Bergee (1997)	Evaluate the accuracy of musical performance by peers, SSAs and teachers.	38	n.s.	Music	Music education	At three locations, college and university voice instructors, percussion, wind instruments, brass and stringed instruments graded for college students. performances. Later, the artists rated the same set of performances on video.	Music skills	Adjudication forms for voice and instruments (scores: 0-100)	Correlations	Correlations between faculty and peer evaluations were generally high. SA correlated poorly with both adult and peer evaluation. No significant differences in SA were found.
Biango-Daniels & Sarvary (2020)	Analysing the impact of experience on SA of their own and their peers' writing accuracy	2,606	n.s.	Sciences	Higher education	Pre-post. Rubric and criteria, SA training and peer-review and rubric provided.	Transferrable skills.	Rubric	Correlations	SA were higher than Peer-assessed. Both self- and peer-assessed grades were significantly lower than the instructor's grade.
Biernat et al. (2003)	To assess the accuracy of self-assessment in the competencies displayed by geriatric residents.	12	n.s.	Medicine	Higher education	Resident medical students assessed a standard patient with symptoms of dementia with a scale. Then the faculty assessors evaluated the patient.	Dementia category	17-item behavioural checklist.	n.s.	Significant difference between SA and functional assessment expert.
Biswas et al. (2015)	Assess the effect of small group sessions on item analysis and SA.	150	17-20	Medicine	Higher education	Three groups are given two pre- and post-tests on the subject knowledge. Between the pre- and post-tests, they work in small groups on the contents. In the post, they also have experience with SA and Feedback.	Carbohydrate chemistry and metabolism	Multiple choice questions (MCQs) 30 items	Correlations	There was statistically significant positive bias or overestimation in SA, both in the pre-test and in the post-test. There is improvement between pre and post in both test score and SA. Gender differences.
Boerebach et al. (2012)	Analysing the relation between surgeons' SA and the residents' evaluation of those surgeons.	663	n.s.	Medicine	Higher education	A scale to assess surgeons' ability and knowledge is developed and administered to residents and surgeons.	Domains of surgical teaching performance	SETQ (System for Evaluation Teaching Qualities)	Kendall's r for Rank Correlations	The correlations between surgeons' self and residents' evaluations were low.
Bolivar-Cruz et al. (2018)	Examines the influence of confidence, self- efficacy and the existence of incentives on SA of their oral presentation competence.	201	n.s.	Business	Higher education	A rubric with criteria is provided, and student are asked to estimate the performance. Then a teacher evaluates. For one group the SA had an impact on the grade and for another group it did not.	Oral presentation competence	Rubric, ten assessment criteria three-level scale	Correlations	The existence of rewards is the only variable that has a significant influence on the self-assessment of male students, while those of female students are determined, above all, by their self-efficacy.
Boud et al. (2013)	Evaluate whether students who voluntarily engage in self-assessment improve in their capacity to make those judgements. Effects of SA over time	182	n.s.	Design	Higher education	Student marks were compared with those from tutors to plot changes over time.	Industrial Design, Visual Communication Design, Fashion and Textile Design and Interior Design	Criteria- based assessment system (ReView)	T-test	Overall students' judgements do converge with those of tutors, but that there is considerable variation across achievement levels, with weaker students showing little improvement.
Boud et al. (2015)	Investigate whether curriculum design has an effect SA accuracy, or whether assessment methods may negate their impact.	1,162	n.s.	Business	Higher education	Student marks were compared with those from tutors to plot changes over time.	Ability business	Criteria- based assessment system (ReView)	Paired T-test	The best students overestimate in the first task but this is reduced in the last task. The most accurate students in SA were also those who improved their performance across tasks.

Table 1 (*continued*)

Author (year)	Aim of the study	Sample size	Age (mean)	Area of knowledge	Educational level	Procedure	Self-assessment task	Measurement of task performance	Analysis	Results
Butler et al. (2006)	Understand the role of context in SA among elementary school students; Examining the validity of 2 types of assessments.	151	n.s.	Languages	Primary education	The results of the 2 types of self-assessments (off-task and on-task SA) were compared both with the students' general proficiency test scores and with their teachers' assessments.	Oral performance	Off-task SA and on-task SA, 5-point likert scale	Correlations	If self-assessments are administered in an on-task format, students can self-assess their oral performance more accurately than they can in an off-task format.
Butterworth (2010)	To develop tools to aid the development of SA skills in Nepali doctors	15	27-30	Medicine	Postgraduate	Three SA tasks per month over a 6-month period; one mini-clinical evaluation exercise, one clinical case review and one significant event analysis. SA was compared with mentor assessment.	Medical skills	Mini-CEX tool, SEA	Correlations	Most doctors were able to accurately self-assess in some areas. Feedback from a senior tutor was vital. Process itself helped to develop awareness of key learning issues.
Carroll (2020)	SA accuracy in supported, criteria-based self-assessment in two business discipline courses.	269	n.s.	Business	Higher education	Quantitative study focuses on SA accuracy on criteria in two connected assessment tasks in two Business School.	Business knowledge	REVIEW	Correlations	Initial over-assessment was widespread and student's under-assessing was associated with higher achievement. Inaccurate participants, consistently responded to feedback of initial inaccuracy and improved accuracy.
Cave et al. (2007)	Examining the effect of using standard assessment criteria during communication skills teaching on students' performance.	359	n.s.	Medicine	Higher education	It was divided into 3 groups, criteria available on the website, criteria explained and criteria provided at the time of the evaluation.	Communication skills	OSCE marksheet	Correlations	There was no significant difference in the end-of-year OSCE performance of students who received the three different conditions. There were low but significant correlations between the tutors and the SA.
Chang et al. (2013)	Exploring the reliability and validity of Web-based portfolio SA	72	n.s.	Informatics	Higher education	Pre-post. Students are assessed by means of a portfolio during the course, at the end of which SA is compared with Tutor and Exam	Writing skills	Rubric	Correlations	1) there was a high consistency level between students' two assessment results; 2) the SA and teacher-assessment results were highly consistent; 3) SA and end-of-course examination were highly consistent,
Chur-Hansen (2000)	Aimed to make clear to 1º year undergraduate medical students the expected writing skills required for an essay examination.	128	n.s.	Medicine	Higher education	A practice essay was written by each student for formative assessment. The essay was rated by a tutor and by the student according to well-defined criteria.	Writing skills	12-items evaluation form	Correlations	Correlations between self- and tutor evaluations were quite low.
Cooney et al. (2021)	Determine if plastic surgery resident trainee self-evaluations differ by resident sex.	64	n.s.	Medecine	Postgraduate	Residents select a case performed, assess it according to a 5-point scale and the instructor assesses using the same scale, including corrections where appropriate.	Surgical skills	Operative Entrustability Assessment (OEA)	t test	Residents significantly underrated their performance during PGY1. In PGY2, males tend to overestimate their performance and women to underestimate it.
Das et al. (1998)	SA y PEER accuracy on medical skill	64	n.s.	Medicine	Higher education	Methodology based on Project Based Learning (PBL) and comparison between different types of evaluation.	Medical knowledge	Assessment scale	Correlations	Mean SA were high throughout the module and did not follow any trend form theme one to theme five.

Table 1 (*continued*)

Author (year)	Aim of the study	Sample size	Age (mean)	Area of knowledge	Educational level	Procedure	Self-assessment task	Measurement of task performance	Analysis	Results
Davey (2015)	Determine the correlation (accuracy) between the student SA and mark of the tutor	32	21-22	Sciences	Higher education	The course materials included detailed lecture notes, presented as four progressive modules, six tutorial assignments and, a design project, together with explicit assessment rubrics.	Thermodynamics knowledge	Rubric	Correlations	Overall student SA was therefore about 1.16 times that of the tutor's mark.
De Grez et al. (2012)	Analysing the agreement between professional assessment and self- and peer assessment of oral presentation skills and student perceptions.	57	(18 years)	Business	Higher education	After the first presentation, students participated individually in a computer-based multimedia training programme about oral presentations. Students received feedback on their first presentation. The intervention was spread over nine lesson weeks.	Oral presentation skills	Rubric	T-test and Correlations	SA are, for the most part, higher than the marks given by teachers. The results also reflect a very positive attitude of students towards peer assessment as a relevant source of external feedback.
Dikici (2009)	Investigating digital portfolio assessment in higher art education with the combination of self, peer and instructor ratings.	34	21-28	Art	Higher education	The students were asked to prepare their portfolio in the computer environment. The students were given 4 weeks to prepare the digital portfolios. The digital portfolios were evaluated as to the rubric.	Content, theme, language and conversations	Rubric	Correlations	Lowest correlation values were found between the instructor and the peer, and the highest correlation values were found between self and the peer.
Dominguez et al. (2016)	Compare self-, peer-, external- and instructor-assessments.	97	n.s.	Informatics	Higher education	The experience was conducted at two different universities during two consecutive years. Students developed websites and online tools were employed to organise the different types of assessments.	Computing products	Rubric	Correlations	There is a high-level of consistency across the different kinds of assessments. The assessment experience correlated with learning gains.
Durant (2014)	To investigate the accuracy of students' self-assessments compared to expert assessments.	24	n.s.	Music	Secondary education	Students recorded themselves playing the instrument and then listened to their recording to complete their SSA. Instructors completed the same assessment.	Music skills	Woodwind/ Brass Solo Evaluation Form (WBSEF)	2 Factors ANOVA	Students' ratings were not consistently higher or lower than the experts' ratings. There was a significant difference in the evaluations across time.
Elhadi et al. (2020)	Assessing students' perceptions and experiences of a paediatric skills program and compared their SA with their preceptor.	65	n.s.	Medicine	Higher education	One-week training course. Participants completed questionnaires self-assessing their performance skills, while examiners evaluated procedural skills.	Procedural clinical skills	Self-assessment questionnaire, 4 likert options, Objective Structured Clinical Examination (OSCE).	Wilcoxon's Z	No significant differences were found between students' self-assessment and preceptors' evaluation scores. Students had accuracy SA.
Ellery et al. (2004)	To establish whether involvement of students in the assessment process assisted them in assessing their own academic worth and was beneficial to their learning.	56	n.s.	Sciences	Higher education	Students provided estimates of their grades immediately prior to an assessment and immediately afterwards. They then graded an assessment written by a peer as well as their own. Finally, teacher evaluate the assignments.	Writing skills	Criteria sheet test	Correlations	Comparisons of these estimates with the final lecturer-awarded grade for the assessment revealed poor correlations initially, but these correlations improved during the course of the study.

Table 1 (*continued*)

Author (year)	Aim of the study	Sample size	Age (mean)	Area of knowledge	Educational level	Procedure	Self-assessment task	Measurement of task performance	Analysis	Results
Emam et al. (2016)	Determine if there is a difference between dental SA of the exam compared to oral surgery faculty.	111	n.s.	Odontology	Higher education	In real time randomized manner, faculty assessed each student using a standardized rubric to assign a score of 1 to 4. Students assessed their own performance.	Surgery skills	Rubric, 4 option likert point.	T-test	Students gave themselves significantly higher grades on the exam compared to the faculty.
Ericson et al. (1997)	To evaluate the impact of guidelines to improve student self-assessment and compare self-evaluations with teacher's evaluations	41	25-28	Medecine	Postgraduate	The patient care course is developed. Students and instructors evaluate their performance through the same guide at the end of the course.	Clinical knowledge	Rubric Guidelines	Mann-Whitney U-test.	Great accordance between SSA and instructor evaluations. However, students under-scored their performance more frequently than they over-scored it.
Esfandiari et al. (2013)	Comparing self-assessors, peer-assessors, and teacher assessors, to determine whether they differed in the levels of severity	188	18-29	Languages	Higher education	During this phase of the training, the lead researcher monitored their ratings and explained any unclear points. Following the training session, each self-assessor rated his/her own essay.	Writing skills	6-point analytic scale.	T-test	Teacher assessors were the most severe while self-assessors were the most lenient, although there was a great deal of variability in the levels of severity that assessors within each type exercised.
Eva & Regehr (2011)	Examining the relationship between SA and self-monitoring	141	(19.35 years)	Psychology	Higher education	Students answer the trivial questions and later answer SA test about their performance.	General knowledge	Self-assessment questionnaire	T-test and Correlations	Poor correlations between performance and SA. Participant performance was strongly related to several measures of self-monitoring.
Evans et al. (2007)	Checking if peer-assessment was more reliable than SA when compared with assessment by a trainer.	38	n.s.	Medicine	Postgraduate	Students, postgraduate and trainers used checklist and global rating scales to evaluate surgical skills in removing a mandibular third molar tooth.	Surgery skills	Checklist and global rating scales	T-test	There was no statistically significant difference between peer-assessed and trainer-assessed scores. SA were significantly higher on average than those given in peer assessment.
Fieler (2020)	Analysing the effects of watching a video about a student task on competency SA.	35	n.s.	Nursing	Higher education	Experimental group participated in the physical assessment competency with video recording. Control group self-evaluated with the same rubric but did not have a video recording to review.	Physical assessment skills	Rubric	n.s.	Experimental group tent to be more accurate.
Fitzgerald et al. (2000)	Examining the SA skills of medical students across two task formats: performance-based and cognitive-based.	304	n.s.	Medicine	Higher education	Performance of a comprehensive clinical assessment examination in two formats: a performance task and cognitive task + SA.	Medical skills	Comprehensive Clinical Assessment examination, self-assessment questionnaire.	T-test and Correlations	The student bias and deviation indices were similar on the cognitive and the performance tasks. The correlations also indicated similarity between the two types of tasks.
Fitzgerald et al. (2003)	Comparing actual and estimated examination performance for three classes during their first 3 years of medical school.	500	n.s.	Medicine	Higher education	Students assessed their performance on classroom examinations and objective structured clinical examination stations during 3 years of medical school.	Medical skills	Objective Structured Clinical Examination (OSCE)	T-test and Correlations	SA accuracy were relatively stable over the first 2 years of medical school with a decease occurring in the third year.
Fitzgerald et al. (2018)	Analysing relationships among self, peer, and faculty marking.	86	18-26	Medicine	Higher education	Students completed an assessment on history-taking skills during a simulated patient scenario and the students SA their performance.	Clinical history-taking and communication skills	Rubric, SHARP tool	Correlations	Correlations between self and peer and self and faculty marks were moderate.

Table 1 (*continued*)

Author (year)	Aim of the study	Sample size	Age (mean)	Area of knowledge	Educational level	Procedure	Self-assessment task	Measurement of task performance	Analysis	Results
Frye et al. (1992)	Comparing the accuracy of SA across four examinations during their first 2 years of medical school.	22	n.s.	Medicine	Higher education	Immediately after answering an average of 20, students predicted their performance. The differences between students' predictions and the experts' scores were calculated.	Medical knowledge, accuracy	Self-assessment sheet	Correlations	Accuracy in self-assessment improved from examination 1 to examination 3 (with less overestimation) and dropped on examination 4 (with more underestimation).
Ganni et al. (2018)	Determine the impact of implementing a self-assessment tool on the concordance between self-assessment and expert assessment.	60	n.s.	Medicine	Postgraduate	Participants of both groups were instructed by the expert surgeons on the procedural tasks of the LC. Test group received an additional training session on SSA before they performed the procedure.	Surgical skills	The Competency Assessment Tool (CAT)	MannWhitney U test	The test group showed better accordance between expert- and self-assessment.
Geranmayeh et al. (2020)	Evaluate the effects of verbal and written feedback in clinical midwifery placement on SA	120	n.s.	Medicine	Higher education	Three-group quasi-experimental study	Clinical midwifery skills	Checklist	Correlations	There was significant direct correlation between the SA scores by both the second instructor and students in the control group
Gonzalez-Betancor et al. (2019)	Evaluating SA accuracy, peer and professor assessments, using scoring rubrics.	155	n.s.	Business	Higher education	The experience consisted of giving an oral presentation in pairs. This presentation was assessed with a rubric. Each presentation was assessed by the speakers, their peers and two professors.	Oral presentations	Rubric	Correlations	Self-assessment accuracy is low and related to the student's gender and that even using a scoring rubric, students receiving higher marks from professors are more accurate than students receiving lower marks.
Graddy et al. (2018)	Explored the impact of training that includes SA and the contrast between self-assessment and expert assessment on outpatient care.	46	n.s.	Medecine	Postgraduate	Students are introduced to the intervention. After the intervention, students complete a self-assessment using a checklist. Feedback session where self-evaluations were compared with those of the experts.	Attention to pacient skills	Checklist	Percentages	Structured episodes of direct observation and coaching in the outpatient setting, with a behavior checklist, appear acceptable and useful for internal medicine residents' learning and development.
Grant et al. (2017)	Determine the utility of SA in microsurgical training using a previously validated rating scale.	8	n.s.	Medicine	Postgraduate	Learners completed multiple self-assessments of their technical skills. Simultaneously, preceptors assessed the learners using the same scale.	Surgery skills	University of Western Ontario Microsurgical Acquisition/Assessment instrument	Correlations	There was a significant agreement noted between the preceptor assessments and self-assessments.
Groenendijk et al. (2020)	Evaluating students SA accuracy	784	12-18	Art	Secondary education	Once the students knew the assessment criteria, they used them to evaluate the lessons. The teachers did the same after the lesson.	Art knowledge	Rubric	T-test and Correlations	The mean self-assessment scores of the students hardly differ from the mean scores given by the teachers.
Guest & Riegler (2021)	Comparing the ability of students to self- and peer-evaluate an extended piece of writing.	110	n.s.	Business	Higher Education	Explanation, assigment, SE and PE.	English skills	Self-peer-tutor evaluation grading sheet.	t test	Peer evaluation is more accurate than SSA but shows greater dispersion.
Guest et al. (2017)	Measure the students' ability to accurately self-evaluate the quality of their own work.	78	n.s.	Business	Higher education	The self-evaluation exercise was introduced on two out of class essay assessments, one in the first year and one in the second year.	Microeconomics knowledge	Multiple choice questions test	Correlations	Students were significantly more accurate at self-evaluating the quality of their work in the second year than they had been in the first year.

Table 1 (*continued*)

Author (year)	Aim of the study	Sample size	Age (mean)	Area of knowledge	Educational level	Procedure	Self-assessment task	Measurement of task performance	Analysis	Results
Hacker et al. (2000)	Examine students' ability to predict and postdict test performance in a classroom context.	99	n.s.	Psychology	Higher education	Specific lessons that the students received focused on the benefits of accurate self-assessment. 1 week prior to taking each of the three exams, students were given practice exams that were parallel versions of the actual exams.	Educational psychology	Three multiple-choice exams, SA Percentage of items correct question	Correlations	High-performing students were accurate, with accuracy improving over multiple exams. Low-performing students showed moderate prediction accuracy but good postdiction accuracy.
Hall et al. (2016)	To investigate to what extent SA of English-Chinese interpretation are accurate and how the accuracy level would change over time.	136	n.s.	Medicine	Higher education	Two peer teaching sessions led by senior medical students in anatomy. They had to self-assess their knowledge about anatomy and complete a 10-question multiple-choice test of content.	Anatomy knowledge	Ten item multiple choice question examination, five-point Likert style	Correlations	Self-assessments of perceived level of knowledge correlated weakly to their performance in their respective objective knowledge assessments. Correlation improved when students rate their performance.
Han et al. (2018)	Investigate to what extent SA of English-Chinese interpretation are accurate and how the accuracy level would change over time.	38	(21 years)	Languages	Higher education	Longitudinal. Over the 10 weeks, the students took part in three performance assessments organised in Weeks. Provide feedback. In each assessment, the students performed consecutive interpreting in three tasks for each interpreting direction.	English-to-Chinese interpretation	Three descriptor-based rating scales.	G coefficient (ρ2)	SA accuracy improved over time for both interpreting directions. The students tended to over-score.
Han (2018)	Investigating the validity of self and peer ratings on three performance dimensions of English-Chinese.	41	(21 years)	Languages	Higher education	For each formative assessment, students had to interpret one speech in English and one speech in Chinese. Two tasks were used in each formative assessment.	English-to-Chinese interpretation	Three descriptor-based rating scales.	MANOVA and Correlations	Students as a group were unable to replicate teachers' ratings, they were able to rank-order their performances in a fairly accurate manner and improved their SA accuracy over time.
Harrington et al. (1997)	To understand how SA influences surgery in relation to other important skills in orthopaedics.	25	n.s.	Medicine	Postgraduate	Surgery residents performed a self-assessment task for ten skills using a new relative ranking method. Supervising faculty assessed residents using the same instrument.	Surgery skills	University of Toronto In-Training Evaluation Report (ITER)	Correlations	The mean correlation between resident and faculty rankings was low, but was higher for junior residents than for senior residents.
Hawkins et al. (2012)	Examining the impact of video recording on the accuracy of SA of basic surgical skills in a UK setting.	31	n.s.	Medicine	Higher education	Final year medical students were videotaped performing a standardised suturing task in a simulated environment. Students were then shown a videotaped. A SA task was then completed.	Surgery skills	Modified validated GRS	Correlations	SA before video feedback demonstrated a moderate positive correlation with expert rates' scores with no change after video feedback. After video feedback, self-assessment scores showed a very strong positive correlation with expert scores.
Herrera-Almario et al. (2016)	Determine the effect of video review on resident and attending assessments of a resident's laparoscopic surgical performance.	9	n.s.	Medecine	Postgraduate	Residents participate in a 7-week Minimally Invasive Surgery rotation operating, Residents received instruction about the GOAL tool before the first surgical procedure. Fill in SSA after done the surgery and after watched the video.	Surgical skills	GOALS tool	T test	Residents overestimate their performance in relation to the experts' assessment.

Table 1 (*continued*)

Author (year)	Aim of the study	Sample size	Age (mean)	Area of knowledge	Educational level	Procedure	Self-assessment task	Measurement of task performance	Analysis	Results
Hewitt (2002)	Determine if a relationship exists between SA accuracy and music performance achievement.	41	n.s.	Music	Secondary education	A pretest-posttest repeated measures design was used in the study. Random sample was used to assign students to one of two treatment groups consisting of either the presence or absence of an aural model.	Music performance achievement	Solo Evaluation section of the Saunders and Holahan (1997) Woodwind Brass Solo Evaluation Form	Correlations	For all subareas, students rated their performances quite positively during each week of the study, with scores ranging from moderate at the beginning of the study to very high in the final.
Hewitt (2005)	Whether differences exist between grade level and music performance subarea (tone, intonation, melody, etc.) on self-evaluation accuracy.	92	n.s.	Music	Secondary education	Two summer music programs self-evaluated their performances during rehearsals, while expert evaluators judged an individual final performance.	Music skills	Solo Evaluation section of the Saunders and Holahan (1997) Woodwind Brass Solo Evaluation Form	Correlations	Results indicated differences between grade levels on performance self-evaluation as the week progressed for some subareas.
Hewitt (2011)	Determine whether SA instruction had an impact on SA, music performance, and SA accuracy of music performance.	211	n.s.	Music	Secondary education	Students in intact classes, grades 5 through 8, were assigned to one of three treatment groups: Participants in the SE-I group received instruction in self-evaluation while students in the SE-O group self-evaluated their performances daily and the SE-No group received no additional instruction.	Music skills	Solo Evaluation section of the Saunders and Holahan (1997) Woodwind Brass Solo Evaluation Form	Correlations	Results suggest that instruction in self-evaluation had little impact on students' self-evaluation accuracy or music performance, although grade level did influence music performance.
Hitzeman et al. (2020)	Enhance understanding of the relationships between postgraduate clinical training, reflective practice, and trainee SA.	37	n.s.	Psychology	Higher education	Mid- and end-of-placement summative assessments of a trainee's practitioner competence using the CΨPRS. Trainees were also asked by each university to evaluate their end-of-placement competencies on the CΨPRS.	Clinical Psychology competences	Clinical psychology practicum competencies rating scale (CΨPRS)	T-test	Trainees are reasonably accurate in their self-evaluations, although they tended to underrate their performance.
Hosein et al. (2018)	Does SA accuracy help metacognition? Investigate factors that could affect.	63	n.s.	Mathematics	Higher education	A few days after taking an exam, the students issued SA and then received the expert's evaluation.	Mathematics skills	Multiple-choice question (MCQ) test	Correlations	Students' accuracy of their self-assessment was found to be associated with their prior mathematical attainment and their overall mathematics confidence.
Hu et al. (2013)	Comparing a video self-assessment of suturing by novice trainees to the assessment by a senior attending surgeon.	23	n.s.	Medicine	Higher education	Medical students were video-recorded while performing suturing. Video SA was carried out within 4 weeks. Both a Global score and total combined OSATS scores were analysed.	Medical skills	Objective structured assessment of technical skills (OSATS).	Correlations	The self-assessment score was significantly higher than a senior trainer's assessment for three tasks by overall score and all five tasks by combined OSATS score.
Hu et al. (2015)	Assess the accuracy of SA among medical students learning basic surgical suturing.	29	n.s.	Medicine	Higher education	Over 7weekly practice sessions, students performed suturing task. Following each task repetition, self- and trainer-assessments (SA–TA) were performed.	Medical skills	36-point weighted checklist	Correlations	Self-assessments tended to overestimate proficiency during the first tercile of practice attempts. Agreement between SA and TA improved with experience.

Table 1 (*continued*)

Author (year)	Aim of the study	Sample size	Age (mean)	Area of knowledge	Educational level	Procedure	Self-assessment task	Measurement of task performance	Analysis	Results
Huth et al. (2017)	Investigate whether agreement between student- and faculty-based assessments increased when students use the assessment criteria.	42	n.s.	Odontology	Higher education	Assessment criteria were available for 6 tasks. Three different assessment training groups. Students' self-assessments of practical tests were compared.	Ontology knowledge	Criteria lists	Correlations and ANOVA	SA showed either consent, overestimation or underestimation. Ratings exhibited significant differences amongst tasks.
Iglesias Perez et al. (2020)	Compare the formative evaluation from the lecturer with the self and peer assessments through a virtual learning environment.	31	n.s.	Social Sciences	Higher education	Students were instructed in the use of the rubric and Moodle. The assignments were randomly assigned and the assessments were carried out.	Statistics knowledge	Rubric	T-test and Correlations	Strong concordance between peer and lecturer assignment, and a moderate agreement between self-assessment and lecturer assignment.
Iguchi et al. (2020)	Student potential for self-assessment in a clinical dentistry practical training course	124	n.s.	Odontology	Higher education	Participants were asked to complete a self-evaluation sheet at the end of each unit of the course	Communication skills	Self-evaluation sheet	Mann-Whitney U	Females tended to rate themselves significantly higher than males
Jahan et al. (2011)	Compare Year 2 students' self-assessment of clinical skills with examiners' assessment of performance in an OSCE using a standard rating scale.	93	n.s.	Medecine	Higher education	Students carry out the OSCE circuit stations. Afterwards they self-assess themselves and the trainers evaluate them with the same instrument.	Clinical skills	Rating scale	Spearman	High accordance between SSA and EA in almost areas.
Juhani et al. (2008)	Investigate how reliably psychology students assessed their professional skills and knowledge.	117	n.s.	Psychology	Higher education	Two cases with different situations that might be a psychologist's work setting, were described in a questionnaire. The students were asked to describe how he/she would act as a psychologist in solving these cases.	Psychology knowledge	Self-evaluation questionnaire, 5 likert options	Correlations	The advanced student group got the best results while the group of beginners got fewer good results. The SA of competence correlated significantly with the number of earned credits in psychology studies and points given by the teacher.
Kachingwe et al. (2015)	To explore the use of videotaping during student physical therapist practical examinations as a mechanism to promote self-assessmen	51	n.s.	Education	Higher education	Groups were asked to self-assess clinical skill and professional behavior immediately following both a midterm and a final practical examination	Physical Therapist Education	Clinical Skills	ANOVA	There was a significant improvement in accuracy regarding the professional behavior score from the midterm to the final
Karnilowicz (2012)	Compare SA amongst undergraduate psychology students and tutor assessments.	64	(21 years)	Psychology	Higher education	Assessment task, SA students and comparison with teacher evaluations.	History of psychology knowledge	Script	Correlations	Students were able to assess their own performance reasonably and accurately.
Keane et al. (2018)	Whether children's SA became more accurate in line with increased age and higher prior literacy attainment.	85	n.s.	Languages	Primary education	Rubric training, English essay and then self-assessed according to the rubric.	Literacy	Rubric developed by Andrade et al. (2008)	Correlations	Children's SA held a weak relationship with their actual performance scores. However, children's SA became significantly more accurate in line with increased developmental stages.

Table 1 (*continued*)

Author (year)	Aim of the study	Sample size	Age (mean)	Area of knowledge	Educational level	Procedure	Self-assessment task	Measurement of task performance	Analysis	Results
Kearney et al. (2016)	Provide a proof of concept of a collaborative peer-, self- and lecturer assessment processes.	280	n.s.	Mathematics	Higher education	The students were presented with the idea of using self- and peer assessments to grade their assignment, which was the creation of a lesson plan for primary mathematics.	Mathematics knowledge	Rubric	Correlations	Students, even those with no prior experience in peer- or self-evaluation, in their first year of tertiary study, under the right conditions, are able to accurately judge their own work and make reasonably accurate judgements of the work of their peers.
Kilic (2016)	Finding out the level of agreement among pre-service teachers' self-, peer- and teacher-assessments of presentation performances	15	19	Education	Higher Education	Students had to prepare a presentation on teaching methods during 30 minutes. Then, they had to evaluate themselves according to an instrument. Rest of students for peer assesment and the instructors used the same instrument. The criteria were previously known to students and instructors.	Teaching skills	5 likert assesment form	ANOVA	Accordance between TEA and SSA.
Kostka (1997)	Whether keyboard skills learning, and SA will have an effect on student attitudes and perceptions of skills.	32	n.s.	Music	Higher education	Instruction and then they had to fill in a scale with 7 answer options on the score they had made. The teacher evaluates them according to these criteria.	Music skills	SA Likert-type scale ranging from 1-7	Correlations	SA were strongly correlated to their posttest perceptions of "knowing" and that knowledge and valuing became more closely associated with instruction and SA procedures.
Kostons et al. (2010)	Investigated the differences in SA and task-selection processes between effective and ineffective learners.	32	(15.9 years)	Sciences	Secondary education	Participants received the pre-test and post-test. The tasks had a multiple-choice format with three answer options and it was possible for multiple options to be correct.	Biology knowledge	Self-assessment questionnaire	Correlations	Although effective learners could more accurately assess their own performance than ineffective learners, they used the same task aspects to select learning tasks.
Kun (2016)	Analyses the SA behaviour and efficiency of undergraduate business students from Hungary.	163	n.s.	Business	Higher education	Students received a pre-examination. This was followed by an intervention phase. Finally, they all underwent a post-test assessment.	Business knowledge	Score test	Correlations	High-achieving students are more accurate in their pre- and post-examination self-assessments, and also less likely to overestimate their performance.
Kustritz et al. (2011)	Compare SA students after completion of internal-medicine clinical rotation.	100	n.s.	Veterinary	Higher education	Students evaluate their clinical competence using the same rotation evaluation form used by supervising teachers. Scores were compared between the two groups.	Clinical competency, knowledge and professionalism	Rubric	Correlations	Students low-performing were more likely to overestimate their competence than were students identified by faculty as higher performing.
Kwasnik & Carter (1999)	This study was undertaken to assess the ability of residents to evaluate themselves	29	n.s.	Medicine	Higher education	Residents were asked to grade themselves in performance characteristics by completing the same forms used by the faculty	Surgical skills	Categorical surgical residents scale	Correlations	A significant correlation was identified between faculty and resident in the global score
Langan et al. (2008)	Examine the effects of gender and level of attainment on the triangulation of marks awarded to students.	60	n.s.	Sciences	Higher education	On the last day of each course, Students gave five-minute presentations, evaluated by tutors, a subset of peers and themselves.	Sciences knowledge	Self-assessment form	Correlations	The ratings generated by peer assessment were more strongly associated with tutor ratings than those from self-assessment.

Table 1 (*continued*)

Author (year)	Aim of the study	Sample size	Age (mean)	Area of knowledge	Educational level	Procedure	Self-assessment task	Measurement of task performance	Analysis	Results
Lappin-Fortin et al. (2014)	Investigate the relationships between SA and experts' assessments in a university.	29	n.s.	Languages	Higher education	Pre-/posttest design. Students submitted their unrehearsed reading from a pretest, which is intended to be a diagnostic learning tool. 12-week course and another reading test was the final grade.	Language skills	Self-assessment form	Correlations	Students were relatively accurate when making a global assessment (Time 1) and when judging some specific aspects of their French pronunciation (Time 2), although they tended to overestimate the extent to which their abilities were native-like.
Lau et al. (2007)	To analyse the reliability and validity of SA of their clinical skills.	9	n.s.	Pharmacy	Higher education	Pharmacists' performance ratings from self, physician, and standardized patient evaluations were compared using Global Rating Scales (GRS) scores and station-specific key points checklists.	Clinical skills	Global Rating Scales (GRS)	T-test and Correlations	Agreement between pharmacists' and patients' GRS ratings ranged from moderate to good.
Lavrysh (2016)	Analysing SA, peer review and teacher assessment	40	n.s.	Engineering	Higher education	The methodology to investigate the effect of peer and self-assessment as a part of the learning process includes literature observation, case study, developing marking criteria, examples of peer and self-assessment strategies and activities.	Language skills	Rubric	Correlations	There were differences between self and peer assessments. Peer assessment was found to resemble more closely teacher assessments whereas self-assessment demonstrated difference with teacher's marks.
Leach (2012)	Assessing SA capacity, peers and the contras with teacher assessments	472	n.s.	Education	Higher education	Each course has an assessment rubric which includes an opportunity for students to self or peer assess.	n.s.	Rubric	Correlations	A statistical analysis showed there was no significant difference between the self-assessed and teacher-assessed grades.
León et al. (2021)	Examines the validity of Student Self-Assessment as an educational assessment in higher education.	74	M=22,48	Education	Higher education	One empirical study is presented that compares student-self evaluations on a test with the evaluation made by the course instructor	Social Education knowledge	Six open-ended questions	T-test and Regresions	Results show a strong correlation overall between the SSA and the instructor's evaluation
Lew et al. (2010)	Assessing SA accuracy and their comparison with peer's assessment and teacher assessments	3,588	(18 years)	Polytechnic	Higher education	The accuracy of the SA of first-year students was studied throughout a semester during which each student made approximately 80 SA. These SA were then compared with peers and tutors.	Students' performance within their team.	Rubric	Correlations	The overall correlations between the scores of self-, peer and tutor assessments suggest weak to moderate accuracy of student self-assessment ability.
Lind et al. (2002)	Evaluate the ability of medical students to perform self-assessment during a third-year surgery	68	n.s.	Medicine	Higher education	Students compared perceptions of their performance with a faculty member's assessment	Medical competencies	11-item, compe-tency-based evaluation form	T test	Female students tend to underestimate their midclerkship performance compared with male
Lindblom-ylänne et al. (2006)	This study focuses on comparing the results of self-, peer and teacher-assessment of student essays	15	n.s.	Law	Higher education	Three people graded each critical essay. First, the student graded the essay. Second, the student graded an essay of peers. Third, the teacher graded all essays.	Knowledge of History of Law	Matrix	n.s.	The comparisons showed that there were fewer differences among self-, peer- and teacher-assessment in the three more technical criteria.

Table 1 (*continued*)

Author (year)	Aim of the study	Sample size	Age (mean)	Area of knowledge	Educational level	Procedure	Self-assessment task	Measurement of task performance	Analysis	Results
Liu et al. (2021)	Evaluate student self-evaluation and compare it with the teacher's self-evaluation.	69	average age = 26. 49 years, range = 24–34 years.	Medecine	Postgraduate	Students were asked to assess their extraction and maxillofacial skills before starting the training. Experts evaluated during the process. Scores were compared after the course.	Surgical skills	DOPS checklist	paired t-testing	Faculty assessments scored significantly higher than did trainees' self-assessments, with no difference between both sexes.
Lopez-Pastor et al. (2012)	To compare the reliability between the SA, the grade awarded by the teacher and the final negotiated grade.	187	n.s.	Education	Higher education	Experimental design. Three groups. At the beginning of the course, the evaluation scales were provided. One group had experience in SA and the other two did not.	Didactics of physical education	Portfolio, Scale with pre-established criteria	Correlations	The results show that there was a high correlation of reliability between assessment in all three programmes.
Lundquist et al. (2013)	Compare SA of their communication skills with the formal assessment of teachers.	401	n.s.	Pharmacy	Higher education	Longitudinal. After an individual oral assessment and again after a group oral assessment, students self-assessed their communication skills.	Communications skills	Rubric	T-test and Correlations	The students' teachers' evaluation scores for both individual and group oral assessments were significantly higher than the students' self-assessment scores.
Marks et al. (2018)	Analyse consistency between supervisors' and interns' self-evaluations based on interns' gender, time of completion.	806	n.s.	Business	Higher education	Upon completion of an internship in the CBE, both supervisors and students perform an evaluation/self-evaluation.	Business skills	5-point Likert-type scale	Correlations	Students, in general, tended to have lower ratings for computer skills, relative to supervisors, suggesting that they do not appreciate the skills that they possess.
Martínez et al. (2020)	Analyse if the e-self-assessment improves student performance in virtual platform Moodle	316	n.s.	Education	Higher education	Student had to fill in 100 question self-assessment questionnaires and a final exam.	Attention to Diversity knowledge	100 question self-assessment questionnaires, Final exam	Correlations	Self-assessment showed improvement of student achievement and increased the degree of student satisfaction.
Martins Machado et al. (2008)	Compare the different perspectives (and ratings) of assessors during tutorials with first-year medical students.	349	n.s.	Medicine	Higher education	Longitudinal. SA, tutor assessment, and peer assessment. We compared these three grades from each tutorial for seven semesters.	Medical knowledge, solve -problem skills	Assessment scale	ANOVA	The SA and Peer were consistently greater than the Tutor assessment. Moreover, the SAA and Per groups did not show statistical difference in any semester evaluated.
Máté & Darabos (2017)	Whether high-achieving students are more accurate in their self-assessment when predicting and evaluating their knowledge	135	n.s.	Business	Higher education	A pretest and posttest design was conducted with a traditional group and a digital group.	Economic knowledge	Traciciones test and Moodle	T-test and Regresions	Higher-achieving students seem to predict and evaluate their examination results more accurately
Máté et al. (2017)	Self-assessment as they predict and evaluate their own performance relative to their externally assessed achievement	159	n.s.	Business	Higher education	Pre and post examination predictions the higher achieving students evaluate their examination results more accurately than their lower.	Enterprise Resource Planning (ERP)	SAP® Software	Linear regression models	Students seem to overestimate their own examination performance
Matsuno (2009)	How do self-assessors, peer-assessors, and teacher-assessors differ when assessing writers' abilities?	91	19-21	Languages	Higher education	Participants received instruction. Make the essay. Evaluation of essays is done in class according to criteria. The students were then instructed to evaluate their own essay and the essays written by five peers.	English-as-a-Foreign Language (EFL) writing	6-point Likert Scale	ANOVA	Many self-raters assessed their own writing lower than predicted. This was particularly true for high-achieving students. Peer-raters were the most lenient ratters; however, they rated high-achieving writers lower and low-achieving writers higher.

Table 1 (*continued*)

Author (year)	Aim of the study	Sample size	Age (mean)	Area of knowledge	Educational level	Procedure	Self-assessment task	Measurement of task performance	Analysis	Results
McKevitt (2016)	Investigate how student engagement with criteria, exemplars, self-assessment, and feedback influenced students' performance.	35	n.s.	Humanities	Higher education	Mixed. 13 weeks. Tutors and students used the same rubric at both draft and final stage to assess performance on the assignment.	Knowledge relating to the presentation essay	Rubric, 5 likert options	Correlations	Overall students' performance in the assignment significantly improved between draft and final submissions. Students' assessment of their work significantly differed to the tutor's.
Miller & Geraci (2011)	Evaluate how college students could improve their ability to accurately predict their own exam performance across multiple exams.	211	n.s.	Psychology	Under-graduates	In two experiments, the degree of feedback received and the level of incentives were manipulated between groups to see how this would affect the accuracy of SA.	Psychology knowledge	Exam	Contrast of % Agreement	Students' predictions were almost always higher than the earned grade, especially low-performers
Muñoz et al. (2007)	Comparing students' oral SA with those of teachers' and inquiring as to students' attitude toward SA.	94	18-20	Languages	Higher education	Prior to the study, teachers were trained. They train their students on how to self-assess they would conduct the actual SA. An attitude questionnaire was given to students.	Language skills	Self-assessment scale, attitudes scale	Correlations	Results of the study showed from moderate to high correlations between teachers' and students' self-evaluations and positive attitudes toward self-assessment.
Nederhand et al. (2020)	Examined how to increase students' awareness of the accuracy of their grade estimates in order to improve their calibration accuracy	261	12-17	Languages	Secondary education	Longitudinal quasi-experimental study. Students provided grade estimates after each of their French exams. Subsequently, the level of reflection support on their earlier estimates was manipulated.	Language skills	Estimate, grade and reflection forms, exam	Regressions	Students in secondary education can learn to provide more accurate estimates of their own exam performance by a simple and easily implemented intervention.
Nejad et al. (2019)	Investigated the effectiveness of self-, peer-, and teacher assessments in assessing EFL students' oral presentations.	60	n.s.	Languages	Foreign Languages education	Mixed. SA before and after the task and peer assessment. Comparison of SAs, peers and teachers. Teachers explain the rating scale criteria for SAs and peer assessments.	Language skills	Rubric	T-test	No significant differences in the three assessment methods, the analysis of the mean scores revealed that teachers employed the strictest scoring criteria while peer assessors used the laxest ones.
Nisly et al. (2020)	Explore whether metacognition can be improved through routine self-assessment	1713	n.s.	Pharmacy	Higher education	Students perform different tasks. In the middle and at the end of each activity, students and preceptors evaluate the performance.	Pharmacy knowledge	Rubric, Exam	Correlations	Students overestimated their performance during the first, second and third trimesters. But with a significant improvement over time.
Oh et al. (2018)	To investigate the effect of teacher calibration and SA on student performance.	165	n.s.	Odontology	Higher education	Instruction + rubric is presented and taught in a lecture + SA is done and handed out + faculty members provide comments.	Odontology skills	Rubric	Correlations	The pilot group's mean SA was significantly higher than the faculty assessment. The mean score of the second-year SA was significantly lower than the faculty assessment.
Ohrstedt & Lindfors (2019)	Analyse whether students taking their first examination were skilled at correctly predicting their achievement.	189	18-46	Psychology	Higher education	Attend course + SA + comparing with teacher assessment	Psychology knowledge	Approaches and Study Skills Inventory for Students. SA scale	T-test	18 % of the students provided perfect ratings while most underestimated their grades. Students reporting the best SA skills expected high grades, but achieved low grades.
Oren (2012)	Examine the relationship of the experience with self/peer assessment with the scores they obtained.	203	19-24	Education	Higher education	Self-assessments and peer evaluations were conducted during the presentations. This application lasted for one semester.	Science	Self-evaluation questionnaire, 5 likert options	T test	The results of the application showed that female students received significantly higher mean scores than male students in all types of scores in terms of the gender variable.

Table 1 (*continued*)

Author (year)	Aim of the study	Sample size	Age (mean)	Area of knowledge	Educational level	Procedure	Self-assessment task	Measurement of task performance	Analysis	Results
Osborne et al. (2014)	Investigate the validity of a self-assessment in the operating theatre and evaluate learning needs.	25	(29 years)	Medicine	Postgraduate	Postgraduate trainees undertook a SA after formally reflecting on appendectomy surgery. Then undertook external assessment. Feedback is given immediately.	Surgery performance	Checklist	Correlations	There were no significant differences in overall scores, learner satisfaction or learning outcomes between external and self-assessment.
Osterhage et al. (2019)	Assess and improve the calibration of students in a first semester introductory biology course.	541	n.s.	Sciences	Higher education	One assesses calibration based on students' SAs and actual grades. Notions of SA are taught. Up to four sections of the course are taught each semester. SA is integrated into the sessions.	Biology knowledge	SA scale, exam	Correlations	Students were significantly mis calibrated for the first test: their predicted scores were, significantly higher than their actual scores. The lowest performing students had the most inaccurate estimates.
Panadero & Romero (2014)	To analyse how the use or non-use of rubrics affected self-assessments.	218	M=22.17	Education	Higher education	A 2x2 quasi-experimental approach was used. Rubric vs. non-rubric, by two occasions pre and post training.	Designing a conceptual map	SA scale, rubric, exam	Correlations	The rubric group reported higher learning strategies use, performance and accuracy.
Papinczak et al. (2007)	Exploring SA, peer and tutor assessment of tutoring performance.	125	n.s.	Medicine	Higher education	Students analyse a practice problem, formulate hypotheses and engage in self-directed learning to try to understand and explain all aspects of the patient's 'problem'.	Analysis clinical case	Assessment scales, 5 likert point	Correlations	Tutor evaluation scores correlated poorly with self-assessment scores, with students consistently underlining their own performance to a substantial degree.
Partido & Henderson (2021)	Evaluate the training effects of photography and self-assessment on ergonomic knowledge	49	n.s.	Dental	Higher Education	Students are instructed on ergonomic knowledge. They receive instructions on how to complete the self-assessment and fill it in. Test group students fill it in before and after viewing. They also receive feedback. Control group did not view the photographs.	ergonomic knowledge	M-DOPAI	ANOVA of Kappa coefficient	Feedback with photography resulted in improved ergonomic scores.Feedback with photography increased the accuracy of the ergonomic self-assessments.
Partido (2020)	To determine the longitudinal effects of feedback and self-assessment on the accuracy.	32	n.s.	Odontology	Higher education	All participants' pre-training and post-training photographs were evaluated for ergonomic scores by two rates.	Dental hygiene	Modified-Dental Operator Posture Assessment Instrument	ANOVA	The accuracy of self-assessments improved for all students who received the ergonomics training.
Pawluk et al. (2018)	To assess the reliability of first-year pharmacy student evaluations by faculty members compared to a standardised patient.	24	n.s.	Pharmacy	Higher education	Pharmacy students completed four stations focused on communication. During each station, students completed a self-assessment and were evaluated by a faculty member.	Communications skills	Assessment scale	Correlations	Student self-assessments were rated higher than the corresponding teacher and standardised patient assessments.
Peyre et al. (2010)	Measure the accuracy of obstetrical and gynaecology resident SA as compared to faculty.	37	n.s.	Medicine	Higher education	Residents completed four laparoscopic procedures. At the conclusion of the session, residents and faculty completed global rating scale assessments.	Laparoscopic technical skill performance	Self-assessment five-point Likert scale	Correlations	Overall residents SA was lower than faculty evaluation.
Pilotti et al. (2019)	Analyse SA Accuracy	685	n.s.	Psychology	Higher education	Classes taught by the same instructor were randomly assigned to a self-assessment practice condition, or to a control condition.	Specific domain of knowledge	New General Self-Efficacy scale	ANOVA	Students in the practice condition displayed not only greater prediction accuracy, but also greater final test performance than students in the control condition.

Table 1 (*continued*)

Author (year)	Aim of the study	Sample size	Age (mean)	Area of knowledge	Educational level	Procedure	Self-assessment task	Measurement of task performance	Analysis	Results
Piper et al. (2019)	To improve the process of student-centred feedback by including a self-assessment component to an assessment task.	754	n.s.	Nursing	Higher education	A two-phased, mixed methods explanatory sequential approach was used, where students use rubrics to SA their assignments	Patient assessment and management	Rubric	T test	Year One students were closer at estimating their own grade than Year Three students. Students often underestimated their grade to see if the marker would match it or provide them with a higher grade.
Pui et al. (2020)	Investigates a disparity between student-teacher rated scores in a criterion-referenced assessment.	151	n.s.	Engineering	Higher education	Using the rubric as a self-assessment tool, students rated their self-confidence level and performance in terms of the critical thinking skills.	Critical thinking skills	Rubric	T-test and Correlations	Despite students' reports of high confidence levels in the acquisition of skills, weak positive correlations between student and teacher-rated scores.
Reitmeier et al. (2009)	Compare self-assessment formats for the evaluation of an oral communication activity.	35	n.s.	Nutrition	Higher education	Students viewed their presentations and evaluated their work using a grading rubric or a reflection assignment. Peer and instructor evaluations were also completed.	Oral Communication Presentations	80 Oral presentation Self-Evaluation Rubric	Correlations	The average scores of students and instructor were similar for the rubric and reflection assessment methods.
Reuland et al. (2009)	To determine the accuracy of medical SA fluency in Spanish.	102	n.s.	Medicine	Higher education	Using predetermined test categories, we determined the predictive value of SA in predicting the same or higher fluency on the test.	Clinical skills, fluency	Assessment scale	Wilcoxon	The predictive value of self-assessment for having at least that level of fluency was 88%.
Rezler (1989)	Explore if problem-based group tutorials help students to improve communication, and SA.	54	n.s.	Medicine	Higher education	Students were randomly. Then reassigned to another group with a different tutor, until they completed the required six units.	Problem-based skills	Assessment scales	Correlations	Medical students in a tutorial program rated themselves in Year 1 and again in Year 2 on Knowledge.
Ricciotti et al., (2012)	Assess resident teaching skills in the resident-as-teacher program	30	n.s.	Medicine	Higher education	Evaluations from the resident-as-teacher training program were compared to evaluations of resident teaching done by faculty.	Teaching skills	Resident-as Teacher Evaluation Tool	Correlations	Resident-as-teacher evaluations were significantly correlated with faculty and resident evaluations
Ritchie (2016)	Compare self-assessment formats for the evaluation of Powerpoint presentations.	39	n.s.	Sciences	Higher education	Students were recorded while giving two PowerPoint presentations on a topic of their choice. All students also received instructor and peer evaluations.	Competencies for oral presentations	Rubrics	Correlations	Students who completed the self-assessment rubric received higher scores than the other group.
Rudy et al. (2001)	Compare faculty, peer, and self-assessment of interviewing skills during a first-year communication and interviewing course	82	n.s.	Medicine	Higher education	Students'self-assessments were compared with the assessments of peers and faculty.	Clinical skill performance	15-point Likert-type scale	ANOVA.	Students are capable of evaluating their peers but have difficulty accurately assessing their own performance
Sadosty et al. (2011)	Determine the accuracy of EM resident performance self-assessment after a simulation-based encounter	17	n.s.	Medicine	Higher education	Residents evaluated their performance immediately after completing simulated cases, and after reviewing the session's video	Performances in Emergency Medicine	Evaluation tool	% and T-test	High- and low-scoring residents accurately self-assessed 83.9% and 62.2%

Table 1 (*continued*)

Author (year)	Aim of the study	Sample size	Age (mean)	Area of knowledge	Educational level	Procedure	Self-assessment task	Measurement of task performance	Analysis	Results
Salehi & Masoule (2017)	Investigate the reliability and validity of self-assessment and peer assessment of Iranian EFL learners' written and oral production data.	32	n.s.	Languages	Secondary education	Explanation about SSA and Peer Assesment. Students did writting tasks and then assess their production. They had to evaluate their oral performance according to another instrument. Teachers and student observers evaluated during the presentation and the presenter did it after his/her presentation.	English skills	essay rating sheet developed by Esfandiari and Myford (2013); five-point Likert scale rating sheet from Peng (2010).	Pearson	SSA was highly reliable in written production assessment. Accordance between SSA and TA in writting tasks.
San Diego et al. (2014)	Determine the level of agreement between staff and SA of clinical skills in performing tasks.	269	n.s.	Odontology	Higher education	Two studies were conducted: Staff members and students rated the students' performance the rubric assessed students' abilities.	Clinical skills	Rubric, 5 likert options	% agreement, Kappa	Agreement between the students' self-assessment and the staff's assessment was high for three of the five criteria.
Sanderson et al. (2016)	Determine if the use of a video-recorded clinical session affects the accuracy of dental hygiene student self-assessment	28	n.s.	Medicine	Higher education	A repeated-assessment experiment where students and experts evaluated a dental hygiene task.	Clinical skill performance	Assessment rubric	T test	No significant differences were found between the differences in overall scores
Sendlhofer et al. (2019)	To test the method of self- and external assessment as a feedback system	58	n.s.	Nursing	Higher education	Self- versus external assessment using a 15 items checklist.	Surgical skills	15 items checklist.	% agreement	Averaging over all checklist items, surgical and non-surgical wards improved only slightly over time
Spoto-Cannons et al. (2019)	Determine differences between SA and their faculty assessments and if they were modified by gender.	535	n.s.	Medicine	Higher education	Mid-term and final assessment and feedback forms from the first-year Doctoring 1 course were analysed from three academic years.	Clinical skills, professionalism	Knowledge test, Assessment scales.	Wilcoxon	Faculty assessments were higher than students and this was not modified by student gender.
Srikumaran et al. (2019)	This study compares resident self-assessed and faculty milestones scores.	21	n.s.	Medicine	Higher education	Residents completed milestone self-assessments before receiving individual score reports.	Communication skills	Milestone rubric	Correlations	For each resident's first assessment, SA and scores were strongly correlated.
Stahl et al. (2020)	Evaluates the accuracy of resident self-assessment versus faculty assessment	110	n.s.	Medicine	Higher education	Assessment data for 5 surgery EPAs was prospectively collected using a mobile application. Matched assessments were identified and the remainder excluded.	General surgery skills	Entrustable Professional Activity (EPA) assessment	T-test	Residents under-rated their own performance relative to faculty assessments. There was moderate agreement between matched resident and faculty assessment.
Stauffer (2011)	This study seeks to investigate ASL students' accuracy in self assessing their language competency	156	n.s.	Education	Higher education	Students and Instructor assessed their ability by reading the levels of the SCPI Rating Scale	Students' language competency	Sign Communication Proficiency Interview (SCPI)	Correlations	There was a significant and moderate-strong correlation between students' self-ratings and their instructors' ratings
Storjohann et al. (2019)	Evaluate accuracy and student perceptions of an innovative grading process that utilizes both peer- and SA.	143	n.s.	Pharmacy	Higher education	Four SOAP note sessions were evaluated. Students scored their note and a blinded peer's note. The average self-, peer-, and final-grades were compared.	Clinical writing	4-point Likert-scale, Subjective, Objective, Assessment and Plan (SOAP)	ANOVA.	No difference was found between the average self-, peer-, and final-grades for all four sessions.

Table 1

Author (year)	Aim of the study	Sample size	Age (mean)	Area of knowledge	Educational level	Procedure	Self-assessment task	Measurement of task performance	Analysis	Results
Swank (2014)	Assessment accuracy in Counselling Competencies	41	25-57	Medicine	Higher education	Compares student ratings and student self-ratings of counselling competencies at midterm and at the end of a practicum course.	Counselling skills, professional dispositions, behaviours.	Counselling Ccompetences Scale (CCS)	T test	Differences were found between faculty supervisors, doctoral student supervisors, and supervisee self-ratings.
Tousignant et al. (2002)	To assess the accuracy of SAA skills of students in a problem-based learning programme.	70	n.s.	Medicine	Higher education	The accuracy of self-assessment was investigated by the relationship between students' self-assessment and performance.	Oral presentation skills	Self-assessment questionnaire, 5 likert option	Correlations	Accuracy is slightly better when the student self-assesses his/her performance ex-post, but the relationship remains very low.
Trofimovich et al. (2016)	Analise the relationship between self- and other-assessment in second language (L2) speech.	132	(23.9 years)	Languages	Higher education	The narratives were recorded. After completing the task, the speakers were asked to use a scale to indicate how well they performed it.	Language skills	9-point assessment scale	Correlations	Mostly inaccurate self-assessment: low scales overestimated their performance; speakers at the high end of each scale underestimated it.
Tuncer et al. (2015)	The aim of this study was to compare the assessment scores of second- and third-year dental students and the faculty	75	n.s.	Odontology	Higher education	After each step, each item was assessed by faculty members, the student, and another student	Restoration types	Preclinical practical exams	Repeated measures analysis	The results indicated that the initial differences decreased.
Unaldi (2016)	Investigates the potential of SA of foreign language skills.	238	n.s.	Languages	Higher education	Language skills were assessed with an objective placement test, and the results were compared with instructors and SA.	Language skills	Proficiency test developed by Allen (1992).	Regressions	Teacher and self-assessment scores were significantly correlated with each other.
Van Hattum-Janssen et al. (2008)	Compare self, peer and teacher assessment in a competence-oriented course.	n.s.	n.s.	Engineering	Higher education	After a training the students made the presentation. The teacher evaluated the presentations. The students evaluated themselves.	Introduction to Civil Engineering	Assessment scale with script	Correlations	The findings show a positive correlation between students' and teachers' grades.
Wagner et al. (2011)	To determine the reliability and value of peer- and self -reported evaluations in the grading of pharmacy students.	234	n.s.	Pharmacy	Higher education	Mean student peer- and self-reported grades were compared to faculty grades in the advanced pharmacy practice experience and seminar presentation courses.	Oral presentations skills	APPE grading rubric, survey instrument	T-test	Self-reported student grades were lower than the faculty-reported grade overall and for the formal presentation component of the APPE course grading rubric.
Ward et al. (2003)	This study aimed to verify the accuracy of self-assessment for the performance of a surgical task and evaluate if can be improved	26	n.s.	Medicine	Higher education	Both students and experts evaluate the videotaped development with two scales	Laparoscopic skills	global rating scale (GRS) and the operative component rating scale (OCRS).	Correlations	Correlation between students and experts moderate, increased after reviewing
Weiss et al. (2005)	How a medical SA clerkship compares to the final institutional grades.	47	n.s.	Medicine	Higher education	Medical students evaluated themselves. Their assessments were then compared with final scores.	Clinical skills	Penn State College of Medicine evaluation tool	Correlations	Statistically significant inter-rater agreement for written/verbal skills.
Wettergreen et al. (2018)	Compare teacher assessment and SA of third-year students' performance in clinical case discussions.	152	n.s.	Medicine	Higher education	Clinical case discussions are held and a self-assessment rubric is administered to the students, which the supervisors	Clinical skills	Rubrics	Correlations	The pooled teacher and student self-assessments correlated for both the first r = 0.41 and second r = 0.35 clinical case discussions.
Wilson et al. (1993)	Examine SA accuracy of students	306	n.s.	Education	Secondary education	Students had to complete two course performance scales and SA. The students' scores are compared with the teacher's score.	Verbal and Numerical Reasoning	Different proficiency scales, performance and self-assessment scales	Correlations	Students tend to overestimate verbal reasoning and the rest tend to underestimate it.

Wong (2016)	Investigating students' and teachers' perceptions of SA skills in two primary schools in Singapore.	18	n.s.	Education	Primary education	They were taught how to use the self-assessment. SA were randomly selected and compared with an independent teacher assessment panel.	Mathematics knowledge	Scoring rubrics	Correlations	Revealed both differences as well as similarities between the students' and teachers' perceptions of students' self-assessment ability.

Note: SA: Self-Assessment; STD: Expert/Standard evaluation; PEER: Peer Evaluation; n.s.: Not specified

Table 2. Variance-related factors in SA

Author (year)	Assessment criteria	Use of Rubric	Self-assessment experience	Feedback	Content knowledge	Incentive	Formative assessment	Expert evaluator	Self-assessment purpose
Abadel et al. (2013)	No	No	No	No	Yes	No	No	Teacher or medical specialist	Accuracy
Abate et al. (2007)	No	No	Yes	No	Yes	No	No	Director	Accuracy
Abdalla et al. (2021)	Yes	Yes	Yes	Yes	No	Yes	No	Teaching staff	Accuracy
Abeyaratne et al. (2022)	Yes	Yes	No	No	No	No	No	Teacher	Accuracy and factors
Admiraal et al. (2015)	Yes	Yes	No	No	No	No	No	Test	Accuracy
Agrawal et al. (2012)	No	No	No and Yes	Yes	Yes	Yes	Yes	Test	SRL (self-monitoring)
Aiko (2018)	No	No	No	No	No and Yes	No	No	Teacher	Accuracy and factors
Aitken et al. (2018)	Yes	Yes	No	Yes	Yes	Yes	Yes	Teacher	Accuracy
Akkus et al. (2017)	No	No	No and Yes	Yes	Yes	No	No and Yes	Teacher	SA as evaluations tool
Akyuz (2018)	No	No	No	Yes	Yes	No	No	Teacher	SA as evaluations tool
Alameddine et al. (2018)	Yes	No	No	No and Yes	Yes	No	No	Teacher	Accuracy
Albanese et al. (2006)	No	No	No	No	Yes	No	No	Test	Accuracy and factors
Alfakhry et al. (2022)	Yes	Yes	No and Yes	Yes	Yes	No	No	Teacher	Accuracy and factors
Ammentorp et al. (2013)	No	No	No	No	Yes	No	No	Doctoral students	Accuracy
Andoh et al. (2008)	Yes	Yes	No	No	No	No	Yes	Teacher	Accuracy and factors
Aryadoust (2015)	Yes	No	No	Yes	Yes	No	No	Teacher	Accuracy
Ashton (2014)	No	No	No	No	No	No	No	Teacher	Accuracy
Austin et al. (2007)	No	No	No	No	Yes	No	No	Teacher	Accuracy
Baars et al. (2014)	Yes	Yes	No and Yes	No	Yes	No	Yes	Test	Accuracy and factors
Baecher et al. (2013)	Yes	Yes	No	No	Yes	No	No	Teacher	Accuracy and factors
Balch (1992)	Yes	No	Yes	No	Yes	Yes	No	Test	Accuracy and factors
Baleghizadeh et al. (2014)	Yes	Yes	Yes	No	No	No	Yes	Teacher	Accuracy and factors

Author (year)	Assessment criteria	Use of Rubric	Self-assessment experience	Feedback	Content knowledge	Incentive	Formative assessment	Expert evaluator	Self-assessment purpose
Ballantine et al. (2007)	No	No	No	No	No	No	No	Test	Accuracy
Bergee (1993)	No	No	No	No	Yes	No	No	Teacher	Accuracy and factors
Bergee (1997)	No	No	No	No	Yes	No	No	Teacher	Accuracy and factors
Biango-Daniels & Sarvary (2020)	Yes	Yes	No	Yes	No and Yes	No	No	Teacher	Accuracy and factors
Biernat et al. (2003)	Yes	No	No	No	Yes	No	No	Teacher	Accuracy and factors
Biswas et al. (2015)	No	No	No and Yes	No and Yes	No and Yes	No	Yes	Test	Accuracy and factors
Boerebach et al. (2012)	No	No	No	No	Yes	No	No	Test	Accuracy
Bolivar-Cruz et al. (2018)	Yes	Yes	No	No	No	Yes	No	Teacher	Accuracy and factors
Boud et al. (2013)	Yes	Yes	No and Yes	Yes	No and Yes	No	Yes	Teacher	Accuracy and factors
Boud et al. (2015)	Yes	Yes	No and Yes	Yes	No and Yes	No	Yes	Teacher	Accuracy and factors
Butler et al. (2006)	No	No	No and Yes	No	No and Yes	No	No	Teacher	Accuracy and factors
Butterworth (2010)	Yes	No	Yes	Yes	Yes	No	No	Teacher	Accuracy and factors
Carroll (2020)	Yes	Yes	No and Yes	Yes	No and Yes	No	No	Teacher	Accuracy and factors
Cave et al. (2007)	Yes	Yes	No	No	Yes	No	No	Teacher	Accuracy and factors
Chang et al. (2013)	Yes	Yes	Yes	Yes	No	No	No	Teacher	Accuracy
Chur-Hansen (2000)	Yes	Yes	No	No	No	No	Yes	Teacher	Accuracy
Cooney et al. (2021)	Yes	Yes	No	No	Yes	No	No	Teacher	Accuracy and factors
Das et al. (1998)	No	No	No	No	No	No	No	Teacher	Accuracy
Davey (2015)	Yes	Yes	No	No	Yes	Yes	No	Teacher	Accuracy
De Grez et al. (2012)	Yes	Yes	No	No	No	No	No	Teacher	Accuracy
Dikici (2009)	Yes	Yes	No	No	Yes	No	No	Teacher	Accuracy
Dominguez et al. (2016)	Yes	Yes	No	No	No	No	Yes	Teacher	Accuracy
Durant (2014)	No	No	No	No	No	No	No	Judges	Accuracy and factors

Author (year)	Assessment criteria	Use of Rubric	Self-assessment experience	Feedback	Content knowledge	Incentive	Formative assessment	Expert evaluator	Self-assessment purpose
Elhadi et al. (2020)	No	No	No	Yes	Yes	No	No	Teacher	Accuracy
Ellery et al. (2004)	Yes	Yes	Yes	Yes	No	Yes	No	Teacher	Accuracy
Emam et al. (2016)	Yes	Yes	No	No	No	No	No	Teacher	Accuracy
Ericson et al. (1997)	Yes	Yes	No	No	Yes	No	No	Teacher	Accuracy
Esfandiari et al. (2013)	Yes	No	No	No	No	No	No	Assessor	Accuracy and factors
Eva & Regehr (2011)	No	No	No and Yes	No	No	Yes	No	Teacher	Accuracy and factors
Evans et al. (2007)	No	No	No	No	Yes	No	No	Assessor	Accuracy and factors
Fieler (2020)	No	Yes	No	No	Yes	No	No	Teacher	Factors
Fitzgerald et al. (2000)	Yes	No	No	No	Yes	No	No	Teacher	Accuracy and factors
Fitzgerald et al. (2003)	Yes	No	No and Yes	No	No and Yes	No	No	Expert	Accuracy and factors
Fitzgerald et al. (2018)	Yes	Yes	No	No	No	No	No	Expert	Accuracy
Frye et al. (1992)	Yes	Yes	No and Yes	Yes	No and Yes	No	Yes	Expert	Accuracy
Ganni et al. (2018)	No and Yes	Yes	No and Yes	No	Yes	No	No	Teacher	Accuracy and factors
Geranmayeh et al. (2020)	Yes	Yes	No	No and Yes	Yes	No	No	Teacher	Accuracy and factors
Gonzalez-Betancor et al. (2019)	Yes	Yes	No	No	No	Yes	No	Teacher	Accuracy
Graddy et al. (2018)	Yes	Yes	No and Yes	Yes	Yes	No	Yes	Teacher	Accuracy
Grant et al. (2017)	Yes	No	No	No	Yes	No	No	Teacher	Accuracy
Groenendijk et al. (2020)	Yes	Yes	No	No	No	No	No	Teacher	Accuracy, SA as tool
Guest & Riegler (2021)	Yes	Yes	No	No	No	Yes	No	Teacher	Accuracy and factors
Guest et al. (2017)	No and Yes	No	No and Yes	No	No and Yes	Yes	Yes	Teacher	Accuracy
Hacker et al. (2000)	No	No	Yes	Yes	Yes	No	Yes	Teacher	Accuracy
Hall et al. (2016)	No	No	No	Yes	Yes	No	No	Expert	Accuracy
Han et al. (2018)	Yes	Yes	No and Yes	Yes	No	No	No	Expert	Accuracy

Author (year)	Assessment criteria	Use of Rubric	Self-assessment experience	Feedback	Content knowledge	Incentive	Formative assessment	Expert evaluator	Self-assessment purpose
Han (2018)	Yes	Yes	No and Yes	Yes	No	No	Yes	Expert	Accuracy and factors
Harrington et al. (1997)	Yes	No	Yes	No	Yes	No	No	Teacher	Accuracy and factors
Hawkins et al. (2012)	No and Yes	Yes	No and Yes	Yes	Yes	No	No	Expert	Accuracy and factors
Herrera-Almario et al. (2016)	Yes	No	No and Yes	No	Yes	No	No	Teacher	Accuracy and factors
Hewitt (2002)	Yes	Yes	No and Yes	Yes	No and Yes	No	Yes	Expert	Accuracy
Hewitt (2005)	Yes	Yes	No and Yes	Yes	No and Yes	No	No	Teacher	Accuracy
Hewitt (2011)	Yes	Yes	No and Yes	Yes	No and Yes	No	No	Teacher	Accuracy
Hitzeman et al. (2020)	No	No	No	No	Yes	No	No	Teacher	Accuracy
Hosein et al. (2018)	No	No	No	Yes	No	No	No	Teacher	Accuracy and factors
Hu et al. (2013)	Yes	Yes	No	No	No	No	No	Teacher	Accuracy
Hu et al. (2015)	Yes	Yes	No and Yes	Yes	No and Yes	No	No	Teacher	Accuracy
Huth et al. (2017)	Yes	Yes	No and Yes	Yes	Yes	No	Yes	Expert	Accuracy
Iglesias Perez et al. (2020)	Yes	Yes	No	No	No	Yes	No	Teacher	Accuracy
Iguchi et al. (2020)	Yes	Yes	No and Yes	Yes	No	No	No	Teacher	Accuracy and factors
Jahan et al. (2011)	No	No	No	No and Yes	No	No	No	Teacher	Accuracy and factors
Juhani et al. (2008)	Yes	Yes	No	No	No and Yes	No	No	Teacher	Accuracy and factors
Kachingwe et al. (2015)	Yes	Yes	No and Yes	No and Yes	No	No	No	Teacher	Accuracy and factors
Karnilowicz (2012)	Yes	Yes	No	No	No	No	No	Teacher	Accuracy and factors
Keane et al. (2018)	Yes	Yes	Yes	Yes	Yes	No	Yes	Teacher	Accuracy
Kearney et al. (2016)	Yes	Yes	Yes	Yes	No	No	No	Teacher	Accuracy
Kilic (2016)	No and Yes	No	Yes	No	No	No	No	Teacher	Accuracy
Kostka (1997)	Yes	Yes	No and Yes	Yes	Yes	No	No	Teacher	Accuracy and factors
Kostons et al. (2010)	Yes	Yes	No and Yes	No	Yes	No	No	Teacher	Accuracy and factors

Author (year)	Assessment criteria	Use of Rubric	Self-assessment experience	Feedback	Content knowledge	Incentive	Formative assessment	Expert evaluator	Self-assessment purpose
Kun (2016)	No	No	No	No	Yes	No	No	Teacher	Accuracy
Kustritz et al. (2011)	Yes	Yes	No	No	Yes	No	No	Teacher	Accuracy
Kwasnik & Carter (1999)	No	No	No	No	Yes	No	No	Expert	Accuracy
Langan et al. (2008)	No	No	No	No	No	No	No	Assessor	Accuracy
Lappin-Fortin et al. (2014)	No	No	No and Yes	Yes	No and Yes	Yes	Yes	Teacher	Accuracy
Lau et al. (2007)	Yes	No	Yes	No	Yes	No	No	Teacher	Accuracy
Lavrysh (2016)	Yes	Yes	No and Yes	Yes	Yes	No	Yes	Teacher	Accuracy
Leach (2012)	Yes	Yes	No and Yes	No	Yes	No	Yes	Teacher	Accuracy
León et al. (2021)	Yes	Yes	No	No	No	No	No	Teacher	Accuracy and factors
Lew et al. (2010)	No	No	No and Yes	No	No	No	No	Teacher	Accuracy
Lind et al. (2002)	No	No	No	No	No	No	No	Teacher	Accuracy and factors
Lindblom-ylänne et al. (2006)	Yes	Yes	No	No	No	No	No	Teacher	Accuracy
Liu et al. (2021)	Yes	Yes	No	No	Yes	No	No	Teacher	Accuracy and factors
Lopez-Pastor et al. (2012)	Yes	Yes	No and Yes	Yes	Yes	Yes	Yes	Teacher	Accuracy
Lundquist et al. (2013)	No	Yes	No	Yes	Yes	No	Yes	Teacher	Accuracy
Marks et al. (2018)	No	No	No	No	Yes	No	No	Expert	Accuracy
Martínez et al. (2020)	No	No	No	No	No	No	No	Test	Accuracy and factors
Martins Machado et al. (2008)	Yes	Yes	No	Yes	No	Yes	No	Teacher	Accuracy
Máté & Darabos (2017)	No	No	Yes	No	No	No	No	Teacher	Accuracy and factors
Máté et al. (2017)	No	No	No and Yes	No	No	No	No	Teacher	Accuracy and factors
Matsuno (2009)	Yes	Yes	Yes	No	No	Yes	No	Teacher	Accuracy and factors
McKevitt (2016)	Yes	Yes	No	Yes	No	Yes	Yes	Teacher	Accuracy
Miller & Geraci (2011)	No	No	No and Yes	No and Yes	No	No	No	Teacher	Accuracy and factors

Author (year)	Assessment criteria	Use of Rubric	Self-assessment experience	Feedback	Content knowledge	Incentive	Formative assessment	Expert evaluator	Self-assessment purpose
Muñoz et al. (2007)	Yes	Yes	No and Yes	Yes	Yes	No	Yes	Teacher	Accuracy
Nederhand et al. (2020)	Yes	No	Yes	Yes	No	No	Yes	Teacher	Accuracy, formative
Nejad et al. (2019)	Yes	Yes	No and Yes	Yes	Yes	No	No	Teacher	Accuracy
Nisly et al. (2020)	Yes	Yes	Yes	No	Yes	No	No	Teacher	Accuracy
Oh et al. (2018)	Yes	Yes	No	Yes	Yes	No	No	Expert	Accuracy and factors
Ohrstedt & Lindfors (2019)	No	No	No	No	No	No	No	Teacher	Accuracy and factors
Oren (2012)	Yes	Yes	No and Yes	No	Yes	No	No	Teacher	Accuracy and factors
Osborne et al. (2014)	Yes	Yes	Yes	Yes	Yes	No	No	Teacher	Accuracy
Osterhage et al. (2019)	Yes	No	No and Yes	Yes	Yes	No	No	Teacher	Accuracy
Panadero & Romero (2014)	Yes	Yes	Yes	Yes	Yes	No	Yes	Expert	Accuracy and factors
Papinczak et al. (2007)	Yes	Yes	No and Yes	Yes	Yes	No	No	Teacher	Accuracy
Partido & Henderson (2021)	Yes	Yes	No and Yes	No and Yes	No	No	Yes	Teacher	Accuracy and factors
Partido (2020)	Yes	No	Yes	No and Yes	Yes	No	No	Teacher	Accuracy and factors
Pawluk et al. (2018)	No	No	No	Yes	No and Yes	No	No	Expert	Accuracy
Peyre et al. (2010)	No	No	No	Yes	Yes	No	No	Teacher	Accuracy
Pilotti et al. (2019)	No	No	No and Yes	No	No	No	Yes	Teacher	Accuracy, formative
Piper et al. (2019)	Yes	Yes	No	Yes	No and Yes	No	No	Teacher	Accuracy and factors
Pui et al. (2020)	Yes	Yes	No	No	No	No	No	Teacher	Accuracy
Reitmeier et al. (2009)	Yes	Yes	Yes	Yes	Yes	No	Yes	Teacher	Accuracy
Reuland et al. (2009)	No	No	No	No	No	No	No	Teacher	Accuracy
Rezler (1989)	Yes	Yes	No	No	No	No	No	Teacher	Accuracy
Ricciotti et al., (2012)	Yes	Yes	No	Yes	Yes	No	No	Expert	Accuracy and factors
Ritchie (2016)	No and Yes	Yes	No and Yes	Yes	No	No	Yes	Teacher	Accuracy

Author (year)	Assessment criteria	Use of Rubric	Self-assessment experience	Feedback	Content knowledge	Incentive	Formative assessment	Expert evaluator	Self-assessment purpose
Rudy et al. (2001)	No	No	No	Yes	No	No	No	Expert	Accuracy
Sadosty et al. (2011)	Yes	Yes	No and Yes	No and Yes	Yes	No	No	Teacher	Accuracy and factors
Salehi & Masoule (2017)	No and Yes	Yes	No and Yes	No	No	Yes	No	Teacher	Accuracy and factors
San Diego et al. (2014)	Yes	Yes	No	No	Yes	No	No	Teacher	Accuracy
Sanderson et al. (2016)	Yes	Yes	No and Yes	No and Yes	No	No	No	Expert	Accuracy and factors
Sendlhofer et al. (2019)	Yes	No	Yes	Yes	Yes	No	No	Expert	Accuracy
Spoto-Cannons et al. (2019)	No	No	No and Yes	Yes	No and Yes	No	No	Teacher	Accuracy and factors
Srikumaran et al. (2019)	Yes	Yes	No and Yes	Yes	Yes	No	No	Teacher	Accuracy and factors
Stahl et al. (2020)	No	No	No	Yes	Yes	No	No	Assessor	Accuracy
Stauffer (2011)	No	No	No	No	No	No	No	No	Accuracy and factors
Storjohann et al. (2019)	No	No	No and Yes	No	No and Yes	No	No	Teacher	Accuracy and factors
Swank (2014)	No	Yes	No and Yes	No	No	No	Yes	Expert	Accuracy
Tousignant et al. (2002)	Yes	Yes	No and Yes	No	Yes	No	No	Expert	Accuracy
Trofimovich et al. (2016)	No	No	No	No	Yes	No	No	Teacher	Accuracy
Tuncer et al. (2015)	Yes	No	No and Yes	No	No and Yes	No	No	Teacher	Accuracy and factors
Unaldi (2016)	Yes	Yes	No and Yes	No	Yes	No	No	Teacher	Accuracy
Van Hattum-Janssen et al. (2008)	Yes	Yes	No	Yes	No	Yes	No	Teacher	Accuracy
Wagner et al. (2011)	Yes	Yes	No	Yes	Yes	No	No	Teacher	Accuracy
Ward et al. (2003)	Yes	No	No and Yes	No and Yes	Yes	No	Yes	Expert	Accuracy and factors
Weiss et al. (2005)	No	No	No	No	Yes	No	No	Teacher	Accuracy
Wettergreen et al. (2018)	Yes	Yes	Yes	Yes	Yes	No	No	Teacher	Accuracy
Wilson et al. (1993)	No	No	No	No	No and Yes	No	No	Teacher	Accuracy and factors
Wong (2016)	Yes	Yes	Yes	Yes	Yes	No	No	Teacher	Accuracy and factors

Note. SA: Self-Assessment; SRL: Self-Regulated Learning

Table 3. Moderator analyses for g_p effect sizes

Moderator / Sub-group	***g***	**95% CI**	***z***	***p***	***k***	***Q***	***df***	***p***
Assessment criteria						1.980	1	.159
No	0.313	[0.106, 0.521]	2.962	.003	106			
Yes	0.142	[-0.028, 0.311]	1.632	.103	169			
Use of Rubric						0.093	1	.760
No	0.231	[0.017, 0.444]	2.117	.034	130			
Yes	0.185	[-0.014, 0.384]	1.825	.068	145			
SA experience						5.789	1	.016*
No	0.219	[0.074, 0.364]	2.963	.003	163			
Yes	0.177	[0.031, 0.324]	2.370	.018	112			
Feedback						4.242	1	.039*
No	0.267	[0.114, 0.421]	3.414	<.001	150			
Yes	0.083	[-0.101, 0.267]	0.882	.378	125			
Content knowledge						6.135	1	.013*
No	0.227	[0.082, 0.373]	3.069	.002	117			
Yes	0.186	[0.041, 0.331]	2.510	.012	158			
Incentive						1.530	1	.216
No	0.170	[0.015, 0.325]	2.147	.032	252			
Yes	0.435	[0.045, 0.825]	2.186	.029	23			
Formative assessment						100.579	1	<.001***
No	0.174	[0.032, 0.316]	2.405	.016	238			
Yes	0.415	[0.268, 0.563]	5.516	<.001	37			
Field of knowledge						20.814	16	.186
Art	0.119	[-0.750, 0.988]	0.269	.788	10			
Biology	-				0			
Business	0.475	[0.061, 0.889]	2.247	.025	16			
Chemistry	0.965	[-0.273, 2.202]	1.527	.127	8			
Computer	0.482	[-0.411, 1.374]	1.057	.290	2			
Education	-0.141	[-0.558, 0.276]	-0.664	.507	32			
Engineering	1.123	[0.218, 2.028]	2.432	.015	5			
Geography	-				0			
Language	0.541	[0.062, 1.019]	2.216	.027	43			
Mathematics	0.019	[-0.700, 0.738]	0.052	.959	9			

Medicine	0.006	[-0.257, 0.268]	0.041	.967	84			
Music	-0.601	[-1.844, 0.643]	-0.946	.344	5			
Nursing	0.436	[-0.792, 1.665]	0.696	.486	2			
Odontology	0.057	[-0.503, 0.617]	0.199	.842	17			
Pharmacy	0.577	[0.068, 1.086]	2.221	.026	21			
Polytechnic	0.693	[-0.531, 1.918]	1.110	.267	4			
Psychology	0.203	[-0.313, 0.720]	0.772	.440	11			
Sciences	-0.625	[-1.876, 0.625]	-0.980	.327	3			
Veterinary	0.237	[-0.997, 1.472]	0.377	.706	3			
Educational level						7.105	3	.069†
Primary education	0.166	[-0.462, 0.794]	0.517	.605	25			
Secondary education	0.076	[-0.812, 0.963]	0.167	.867	7			
University	0.253	[0.103, 0.403]	3.306	<.001	230			
Postgraduate	-0.750	[-1.478, -0.022]	-2.020	.043	13			

Note: g = effect size. 95% CI = lower limit and upper limit of the 95% CI; z = z-score associated with the g value in the same row; p = p-value associated with the z-score in the same row; k = number of effect sizes contributing to g in the same row; Q = result of the Q-test for moderation; df = degrees of freedom of the Q-test for moderation; p = p-value of the Q-test for moderation; - = Redundant predictors dropped from the model.
† $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

Table 4. Moderator analyses for Fisher's *z effect sizes*

Moderator / Sub-group	***Fisher's z***	**95% CI**	***z***	***p***	***k***	***Q***	***df***	***p***
Assessment criteria						0.583	1	.445
No	0.440	[0.337, 0.544]	8.361	<.001	66			
Yes	0.490	[0.414, 0.565]	12.629	<.001	98			
Use of Rubric						2.556	1	.110
No	0.416	[0.324, 0.509]	8.820	<.001	83			
Yes	0.515	[0.434, 0.596]	12.456	<.001	81			
SA experience						3.308	1	.069†
No	0.456	[0.392, 0.519]	14.077	<.001	113			
Yes	0.512	[0.437, 0.586]	13.504	<.001	51			
Feedback						6.264	1	.012*
No	0.423	[0.350, 0.495]	11.350	<.001	109			
Yes	0.559	[0.467, 0.651]	11.937	<.001	55			
Content knowledge						0.451	1	.502
No	0.483	[0.414, 0.551]	13.836	<.001	71			
Yes	0.463	[0.397, 0.530]	13.653	<.001	93			
Incentive						0.703	1	.402
No	0.481	[0.416, 0.546]	14.515	<.001	145			
Yes	0.421	[0.287, 0.555]	6.163	<.001	19			
Formative assessment						1.148	1	.284
No	0.459	[0.393, 0.525]	13.624	<.001	142			
Yes	0.550	[0.395, 0.706]	6.937	<.001	22			
Field of knowledge						19.268	15	.202
Art	0.566	[0.243, 0.890]	3.428	.001	9			
Biology	0.386	[0.015, 0.757]	2.039	.041	2			
Business	0.407	[0.176, 0.638]	3.456	<.001	11			
Chemistry	-				0			
Computer	0.765	[0.431, 1.099]	4.485	<.001	3			
Education	0.555	[0.375, 0.734]	6.046	<.001	16			
Engineering	0.161	[-0.188, 0.510]	0.907	.365	5			
Geography	0.374	[-0.101, 0.849]	1.542	.123	2			
Language	0.642	[0.479, 0.804]	7.755	<.001	23			
Mathematics	0.867	[0.364, 1.371]	3.377	<.001	1			

Medicine	0.423	[0.329, 0.517]	8.809	<.001	68			
Music	0.384	[-0.388, 1.156]	0.975	.329	4			
Nursing	-				0			
Odontology	0.140	[-0.335, 0.615]	0.579	.563	2			
Pharmacy	0.277	[-0.185, 0.739]	1.174	.240	1			
Polytechnic	0.234	[-0.202, 0.671]	1.052	.293	1			
Psychology	0.455	[0.224, 0.686]	3.860	<.001	12			
Sciences	0.518	[0.154, 0.882]	2.786	.005	4			
Veterinary	-				0			
Educational level						10.274	3	.016*
Primary education	0.791	[0.445, 1.138]	4.473	<.001	5			
Secondary education	0.717	[0.517, 0.917]	7.021	<.001	23			
University	0.440	[0.377, 0.502]	13.809	<.001	131			
Postgraduate	0.382	[0.020, 0.744]	2.067	.039	5			

Note: g = effect size. 95% CI = lower limit and upper limit of the 95% CI; z = z-score associated with the g value in the same row; p = p-value associated with the z-score in the same row; k = number of effect sizes contributing to g in the same row; Q = result of the Q-test for moderation; df = degrees of freedom of the Q-test for moderation; p = p-value of the Q-test for moderation; - = Redundant predictors dropped from the model.
† $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

Figure Captions

Fig1. PRIMA Flowchart.

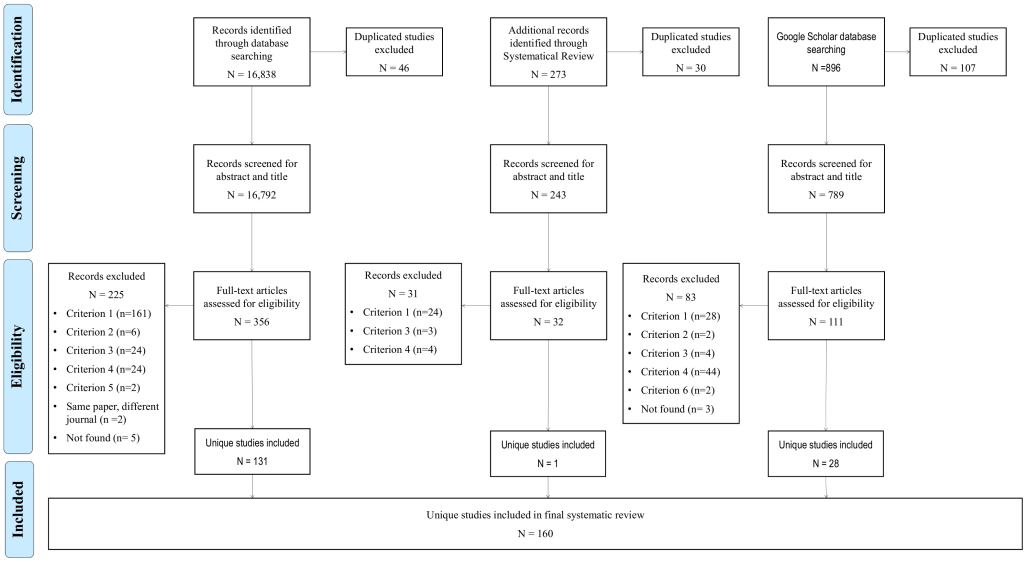


Fig2. Number of publications included in the review by years. The red line represents the developing trend of publications across years.

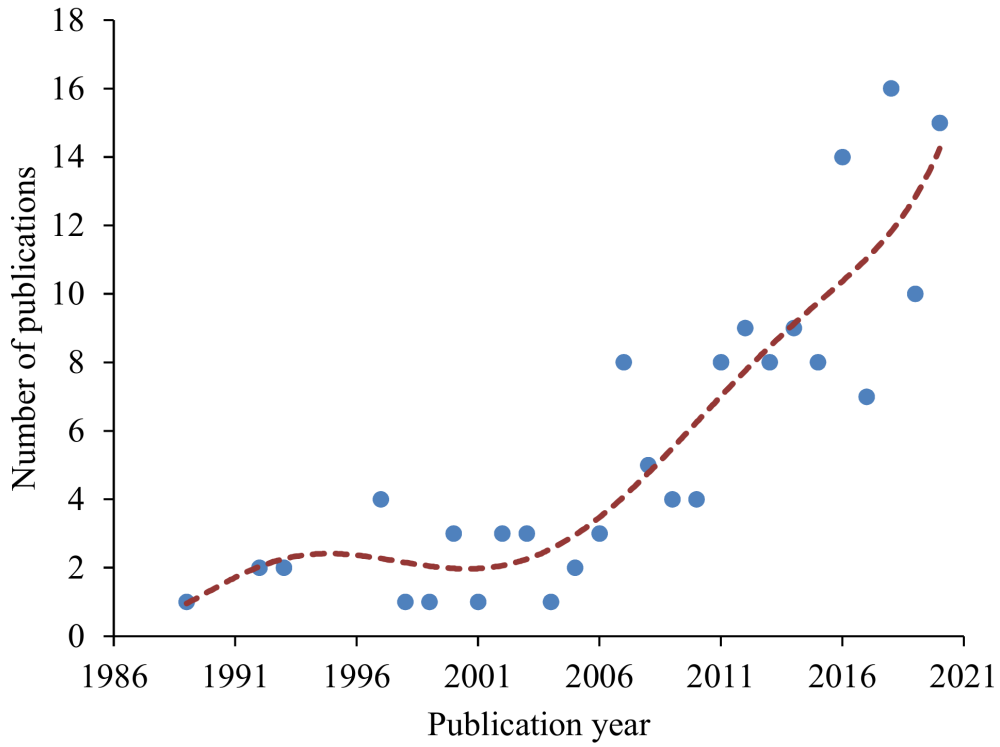


Fig3. Funnel plots for the meta-analysis of g_p scores (panel A) and Fisher's z scores (panel B).

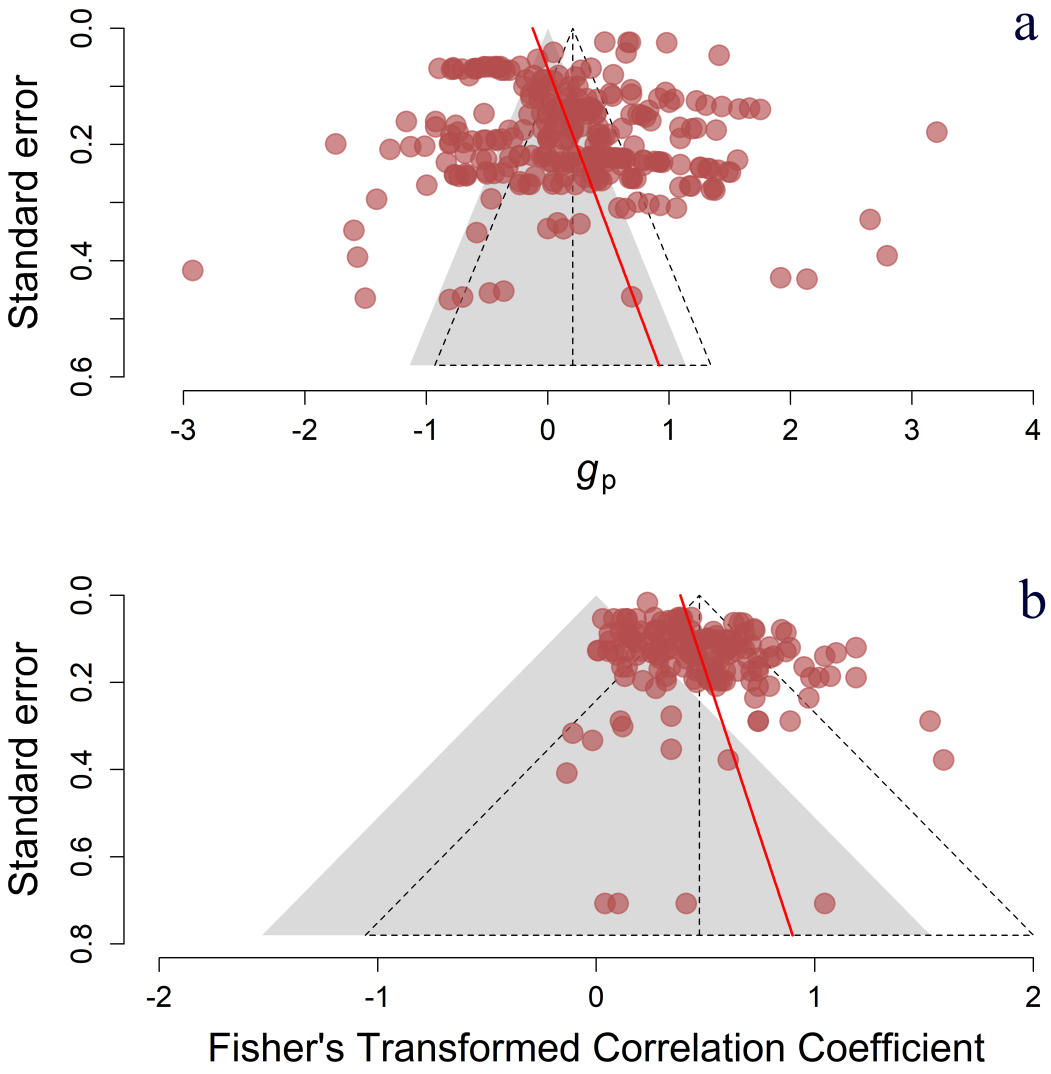
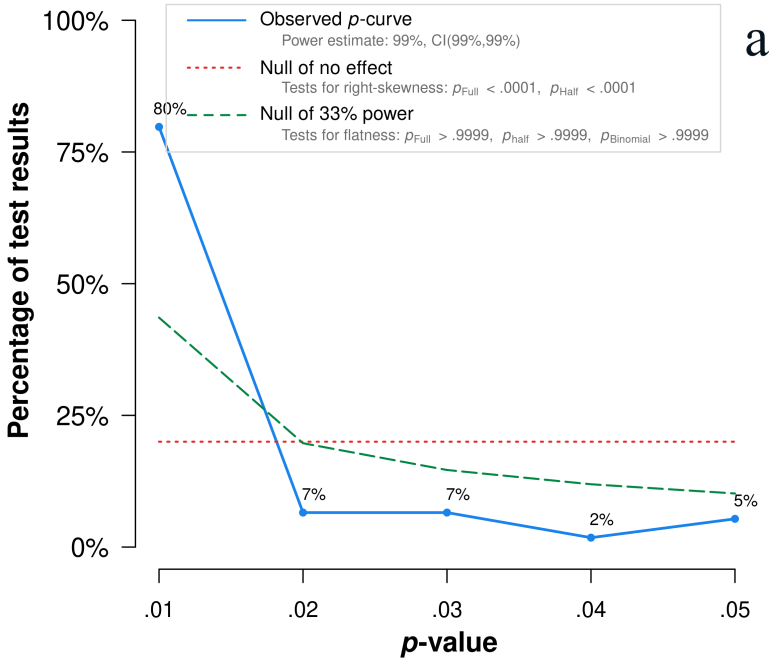
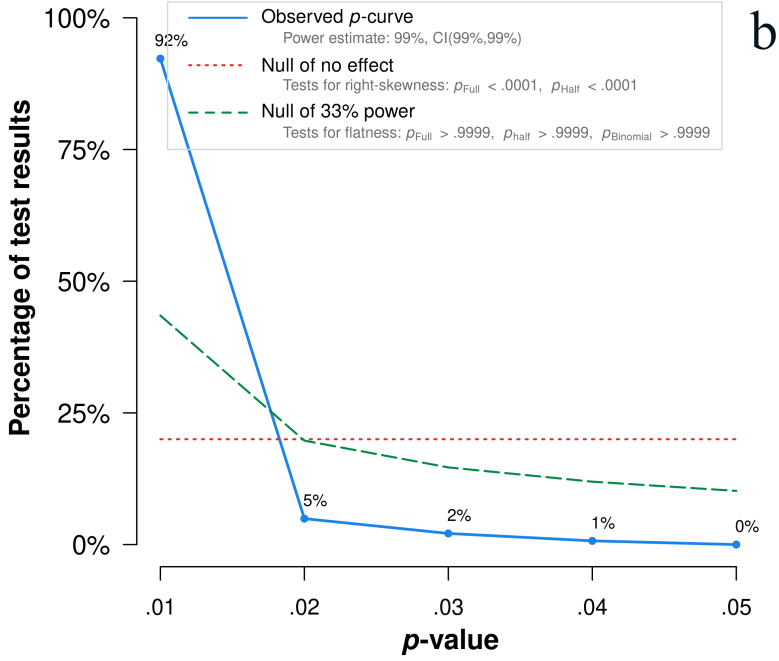


Fig4. *p*-curve plots depicting the distribution of significant *p* values (panel A for *t*'s and panel B for *r*'s).



Note: The observed *p*-curve includes 168 statistically significant ($p < .05$) results, of which 150 are $p < .025$. There were 98 additional results entered but excluded from *p*-curve because they were $p > .05$.



Note: The observed *p*-curve includes 142 statistically significant ($p < .05$) results, of which 140 are $p < .025$. There were 26 additional results entered but excluded from *p*-curve because they were $p > .05$.

Fig5. Forest plot for moderator Meta Regressions. Panel A for g_p and Panel B for Fisher's Correlation.