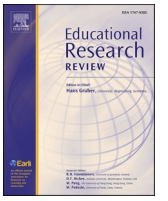
Educational Research Review 37 (2022) 100484

Contents lists available at ScienceDirect

Educational Research Review

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



Review

Effects of self-assessment and peer-assessment interventions on academic performance: A meta-analysis

Zi Yan [a,*], Hongling Lao [a], Ernesto Panadero [b,c], Belen Fernández-Castilla [d], Lan Yang [a], Min Yang [a]

[a] Department of Curriculum and Instruction, The Education University of Hong Kong, Hong Kong, China
[b] Facultad de Psicología y Educación, Universidad de Deusto, Bilbao, Spain
[c] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
[d] Department of Psychology and Educational Sciences, KU Leuven, Belgium

ARTICLE INFO

Keywords:
Self-assessment
Peer-assessment
Meta-analysis
Academic performance

ABSTRACT

This meta-analysis examined the effects of self-assessment (SA) and/or peer-assessment (PA) interventions on academic performance. The synthesis included 626 effect sizes from 175 independent studies, either using an experimental/quasi-experimental design or a repeated measures design, and involved 19,383 participants in total. Results indicated that SA ($g = 0.585$), PA ($g = 0.606$), and SA + PA (mixed) intervention ($g = 0.448$) had meaningful effects on academic performance. The difference between the effects of SA and PA interventions, conducted on different groups within the same study, was not statistically significant. The use of online technology increased the effect of PA interventions but not for SA. Participants with a higher mean age had more performance gains from the SA + PA (mixed) intervention. For both SA and PA, the studies that used a repeated measures design generated larger effect sizes than those with an experimental/quasi-experimental design. Overall, the findings from this meta-analysis demonstrated the benefits of SA and PA interventions on academic performance in different contexts. Implications for practice and directions for future research are discussed.

1. Introduction

Self-assessment (SA) and peer-assessment (PA) have great potential to improve students' academic performance as shown by formative assessment literature (Black & Wiliam, 2009). SA and PA can produce feedback to oneself or others that inform adjustments to learning strategies and enhance performance (Andrade, 2019; Panadero et al., 2018). Much empirical evidence attests to these effects (Brown & Harris, 2013; Huisman et al., 2019; Sanchez et al., 2017).

As SA and PA require students to assume an active and reflective role, understand and apply assessment criteria, seek and use feedback, and evaluate their own or others' work, they are regarded as being effective methods to achieve important educational goals, such as developing self-regulated learning (Andrade & Heritage, 2018; Harris & Brown, 2018; Panadero et al., 2017; Zimmerman & Moylan, 2009), evaluative judgment (Tai et al., 2018), and feedback literacy (Carless & Boud, 2018; Yan & Carless, 2021). Such goals recognise that learning is more than the simple acquisition of knowledge, and involves student agency in judging their own work and

* Corresponding author. Department of Curriculum and Instruction, The Education University of Hong Kong, D1-1/F-49, 10 Lo Ping Road, Tai Po, N.T., Hong Kong, China.
E-mail address: zyan@eduhk.hk (Z. Yan).

https://doi.org/10.1016/j.edurev.2022.100484
Received 26 December 2021; Received in revised form 11 September 2022; Accepted 20 September 2022
Available online 27 September 2022
1747-938X/

proactively seeking and using inputs from others. Thus, SA and PA have received intensive attention in the research literature in higher education contexts (Dochy et al., 1999; Falchikov & Boud, 1989) and K-12 settings (Brown & Harris, 2013; Sanchez et al., 2017).

Usually, SA and PA are studied independently. Researchers have compared the performance of students who receive an SA or PA intervention with those who receive no intervention or different interventions (e.g., Panadero et al., 2014; Yan et al., 2020). However, some studies have used a mixed intervention where SA and PA are combined together (e.g., Lee et al., 2016; Ratminingsih et al., 2018). Other studies have included both SA and PA interventions, even if the two interventions are applied to different groups of students (e.g., Cahyono & Amrina, 2016; van Ginkel et al., 2017). This kind of combined design could enable a direct comparison between the effects of SA and PA interventions if the research procedure is correctly implemented. As studies on this topic accumulate with time, it is both interesting and meaningful to perform a systematic summary of the effects of different designs.

Thus, the present meta-analysis aimed to (a) determine the overall effect of SA and PA interventions on students' academic performance, (b) compare the effects of SA and PA interventions on students' academic performance, and (c) explore the possible moderators influencing the effects of SA and PA interventions.

1.1. What is self-assessment and peer-assessment?

There are multiple definitions for SA and PA as they can be implemented in various forms. In a state-of-the-art review, Panadero et al. (2016) identified 20 categories of SA implementations, varying from a simple form of awarding a grade/mark to one's own work (i.e., self-grading or self-marking) to a more complex form that involves a process in which students seek and use feedback, evaluate their own work against selected criteria to identify their own strengths and weaknesses. Thus, considering those various forms, those scholars defined SA as 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. 2).

Similarly, the varying forms of PA range from simply assigning a grade/mark to peers' work (i.e., peer-grading) to processes whereby students evaluate their peers and offer feedback or are evaluated by peers and receive feedback (Panadero et al., 2018). One of the most extended definitions is that PA is "an arrangement in which individuals consider the amount, level, value, worth, quality or success of the products or outcomes of learning of peers of similar status" (Topping, 1998, p. 250). Despite the variations of SA and PA forms, a general consensus is that SA and PA can reach a higher potential to increase students learning if implemented with formative purposes (Andrade, 2019; Panadero et al., 2018; Yan & Brown, 2017), yet even with summative purposes self and peer grading have shown positive impact in academic performance (Double et al., 2020; Sanchez et al., 2017). While there are important lessons to be extracted from feedback literature, our aim was to focus on what can be concluded from SA and PA research. Therefore, our literature search aimed to explore the effects of those two feedback agents: the self and peers.

1.2. Why is self-assessment and peer-assessment useful in enhancing students' academic performance?

Engaging in SA and PA is an important learning opportunity for students (Yan & Boud, 2021) to develop their self-regulated and co-regulated learning, which has the potential to enhance students' academic performance. SA and PA require students to make judgments about their own and others' work, identify the gap between their current performance and the desired standard and take actions to close the gap (Andrade, 2010; Black & Wiliam, 1998; Butler & Winne, 1995). With a sensible understanding of quality work and a realistic sense of their own performance, students' learning can be more goal-oriented and effective (Baas et al., 2015; Boud et al., 2013). SA is a central component of self-regulated learning and influences each self-regulated learning phase (Panadero, Lipnevich, & Broadbent, 2019; Yan, 2020). During self-regulated learning, students need SA to keep monitoring their learning processes and products against standards and goals (Panadero, 2017), and identify their own strengths and weaknesses aiming for further improvement (Butler & Winne, 1995; Yan & Brown, 2017; Zimmerman & Moylan, 2009). Engaging in PA activities also leads to enhanced self-regulation skills (Meusen-Beekman et al., 2015; Nicol & Macfarlane-Dick, 2006). Furthermore, PA provides opportunities of interacting with others so that students can influence and learn from each other's processes of self-regulation, which allow them to develop internal standards by which to evaluate the quality of their own work and their peers' work. This is also known as co-regulation (Allal, 2016; Panadero, Broadbent, et al., 2019), which serves as a transitional process for students' acquisition of self-regulation skills (Hadwin & Oshige, 2011).

The benefits of SA and PA can also be attributed to their positive impact on learning motivation. SA and PA allow students to take an active and reflective role in the assessment process, resulting in activated student agency, a greater sense of autonomy and ownership (Black & Wiliam, 1998) and reduced dependence on the teacher (Brown & Harris, 2013). Critically engaging in SA and PA is also likely to enhance students' self-efficacy in the tasks being performed (Panadero et al., 2016). According to self-determination theory (Ryan & Deci, 2000), enhanced autonomy and self-efficacy could lead to higher intrinsic motivation to learn.

Nevertheless, SA and PA have very noticeable differences. SA refers to evaluating one's own work and, though the process may involve others (e.g., teacher's feedback), requires a high level of self-awareness, self-reflection and self-judgment. It is an inwards process, though the best instructional strategy to develop SA is trying to make the process as explicit as possible (Panadero, Lipnevich, & Broadbent, 2019). In contrast, PA requires students to evaluate each other's work and they can take up two roles as assessor or assessee. PA involves interpersonal processes, so mutual trust or psychological safety is crucial (Panadero, 2016; van Gennip et al., 2009). According to Vygotsky's (1978) social development theory, PA offers opportunities for student interaction, promoting student learning and development because it provides the look of an "other" from the peer's perspective, which students might easily understand. Logically, SA and PA involve different cognitive, motivational and emotional processes, and for that reason, we need to compare their effects on enhancing academic performance.

1.3. Effects of self-assessment and peer-assessment interventions on academic performance

There is a general consensus that students' academic performance can be enhanced through well-designed SA (Brown & Harris, 2013; Dochy et al., 1999) or PA activities (Dochy & McDowell, 1997; van Zundert et al., 2010). Thus, the implementation of SA and PA has become popular in classrooms across sectors and subjects to enhance student learning. In particular, a large body of research has investigated the effectiveness of SA and PA interventions in improving students' academic performance using experimental/quasi-experimental research with comparison groups (e.g., van Ginkel et al., 2017; Wang et al., 2017; Yan et al., 2020) and repeated measures studies (e.g., Cheng et al., 2015; Noroozi et al., 2016; Pai, 2015; Panadero et al., 2014). Accordingly, the body of research synthesis to evaluate the effectiveness of SA and PA interventions continues to expand. We have summarised the available meta-analyses on this topic in Table 1 and elaborated the main findings in the following sections.

1.3.1. Prior meta-analyses of the effect of self-assessment on academic performance

We found only two meta-analytical reviews of the effectiveness of SA interventions on students' academic performance published in the past ten years. Both reported a generally positive impact of SA, with varied learning gains across studies. Brown and Harris (2013) reviewed 23 studies in K-12 settings. They found that, in general, SA enhanced academic performance across year levels and subject areas (median effect between 0.40 and 0.45). However, the extent of learning gains differed across studies, with some studies reporting nil to small effects (i.e., $d \leq 0.20$). In another meta-analysis in K-12, Youde (2019) reviewed 19 research studies between 1991 and 2017 and reported an overall effect size of 0.46 on academic achievement across year levels and subjects.

1.3.2. Prior meta-analyses of the effect of peer-assessment on academic performance

Relatively, PA interventions on students' academic performance have attracted more attention. Several meta-analyses were published in 2019 and 2020. All of these reviews reported positive effects of PA interventions. Zheng et al. (2020) synthesised 37 studies involving technology-facilitated PA conducted in higher education and K-12 settings, published between 1999 and 2018. An overall mean effect size of 0.576 of technology-facilitated PA on academic performance was found. The use of extra supporting strategies (e.g., extra tools and PA rules) in technology-facilitated PA also significantly affected students' academic performance with an overall mean effect size of 0.543. Huisman and colleagues (2019) synthesised 24 quantitative studies focusing on the effect of formative peer feedback on students' academic writing in higher education. They found that students engaging in peer feedback had large writing improvements compared to students without feedback ($g = 0.91$ [0.41, 1.42]). Peer and teacher feedback's impact was similar in enhancing writing performance ($g = 0.46$ [−0.44, 1.36]). In another meta-analysis, however, Wisniewski et al. (2020) revealed that peer feedback ($d = 0.85$ [0.59, 1.11]) was more effective than teacher feedback ($d = 0.47$ [0.43, 0.51]), although caution was needed as only eight studies about peer feedback were included in the analysis. Two meta-analytical reviews (i.e., Double et al., 2020; Li et al., 2020) provided syntheses on different types of PA across education levels. Double et al.'s (2020) review included 141 effect sizes from 54 studies with experimental/quasi-experimental designs. Results showed that PA had a positive effect on academic performance compared with no assessment ($g = 0.31$, $p < .01$) and teacher assessment ($g = 0.28$, $p < .01$). Li et al. (2020) synthesised 134 effect sizes from 58 studies. They found that PA improved academic performance compared with no assessment ($g = 0.33$, $p < .01$) and teacher assessment ($g = 0.26$, $p < .01$).

1.3.3. Prior meta-analyses covering the effect of both self-assessment and peer-assessment

Meta-analyses covering both SA and PA are scarce. The only relevant one was conducted by Sanchez et al. (2017). They focused on self- and peer-grading, which was defined as "evaluate the extent to which performance criteria and standards have been met (Boud, 1991) and provide criterion-referenced feedback, that is, grading, to themselves or others" (p. 1409). The results based on 33 studies indicated that self- and peer-grading had a positive effect on academic performance. The mean effect size of self-grading was 0.34 ($n = 20$), although 12 of the 44 effect sizes were negative. The mean effect size of peer-grading was 0.29 ($n = 7$). The mean effect size of SA was slightly larger than that of PA. However, they did not report whether the difference was significant or not. Some meta-analyses did SA/PA comparisons based on a small number of studies with both SA and PA groups. For example, Huisman et al. (2019) compared SA and PA intervention effects based on three studies that had both SA and PA conditions (Cahyono & Amrina, 2016; Diab, 2011; Stellmack et al., 2012). The composite effect size of the comparison was small but significant (0.33 [0.01, 0.64]), favouring PA. Li et al. (2020) identified 20 effect sizes comparing SA and PA groups and found no significant difference. Double et al.'s (2020) review had 10 studies with both SA and PA groups. The results also indicated no significant difference between the effect of SA and PA. It appears that the results of the comparison between SA and PA are inconclusive. However, interpretation of the results warrants cautions as these

Table 1
Recent meta-analyses of the effectiveness of SA and PA on academic performance.

Publication Year	Author(s)	Theme	Target population	Number of included studies
2013	Brown and Harris	SA	K-12	23
2019	Youde	SA	K-12	19
2017	Sanchez, Atkinson, Koenka, Moshontz, and Cooper	SA/PA	K-12	33
2019	Huisman, Saab, van den Broek, and van Driel	PA	Higher education	24
2020	Double, McGrane, and Hopfenbeck	PA	Cross sector	54
2020	Li, Xiong, Hunter, Guo, and Tywoniw	PA	Cross sector	58
2020	Zheng, Zhang, and Cui	PA	Cross sector	37

comparisons were based on a small number of effect sizes.

1.3.4. Limitations of prior reviews

The available meta-analytical reviews on SA and PA effect on academic performance have four limitations. Firstly, comprehensive meta-analytical syntheses on the SA effect are scarce. Both Brown and Harris' (2013) and Sanchez et al.'s (2017) reviews are constrained to K-12 settings. The most recent meta-analysis in this area is a Doctor of Education dissertation (Youde, 2019), which covers only 19 studies in K-12 populations conducted between 1991 and 2017, and most of them were dissertations.

Secondly, most of the meta-analyses on the PA effect focus on a specific setting, therefore, limiting their generalisation to other settings. For instance, some focus on one particular PA form (e.g., technology-facilitated PA; Zheng et al., 2020). Some include studies only in higher education (e.g., Huisman et al., 2019) or K-12 (e.g., Sanchez et al., 2017). There are two exceptions (i.e., Double et al., 2020; Li et al., 2020) that provide a synthesis of different types of PA across education levels. However, neither coverage of relevant studies in these two reviews was comprehensive, with 54 and 58 included studies, respectively, and only 28 of them overlapped. The main reason for the relatively small number of included studies was the strict inclusion criteria in both meta-analyses (e.g., experimental or quasi-experimental design only). The small proportion of overlapping studies was probably due to the difference in search timeframe, search keyword, and the definition of outcome variables between these two meta-analyses. For example, Li et al. (2020) searched from 1950 to 2017, whereas Double et al. (2020) searched up to June 2018. Li et al. (2020) used seven synonyms for PA as search keywords, while Double et al. (2020) used four synonyms. Li et al. (2020) focused on cognitive skills only (i.e., either examination scores or performance results), but Double et al. (2020) adopted a broader definition, including both traditional learning outcomes and practical skills (e.g., constructing a circuit).

Thirdly, there are very few meta-analytical reviews covering both SA and PA. We argue that interpreting the results of SA and PA studies together and comparing the effects of SA with PA are meaningful for both theoretical and practical reasons. Theoretically, both SA and PA emphasise students' active role in the assessment process and are pertaining to critical competencies, such as regulation of learning and evaluative judgment (Panadero, Broadbent, et al., 2019). Empirically, SA and PA are often implemented together in practice because they offer the potential for triangulation on learning progress (Sanchez et al., 2017). For example, the SA process described in Yan and Brown's (2017) model explicitly endorses the role of peer feedback in contributing to the SA judgment (see also To & Panadero, 2019). Two reviews (i.e., Dochy et al., 1999; Topping, 2003) covered both SA and PA, but they are narrative reviews and somewhat outdated. Sanchez et al.'s (2017) review on self-grading and peer-grading is constrained to K-12 settings and included only a small number of studies (20 studies on SA and 7 studies on PA), partly because they focused on self- and peer-grading, rather than embracing a broader conception of SA and PA. Furthermore, Sanchez et al. (2017) did not report whether the effect sizes of self-grading and peer-grading were significantly different from each other.

The current review, hence, attempted to address these limitations. First, we included studies examining the effect of SA and/or PA interventions on academic performance so that we could interpret and compare the effect of SA and PA interventions in the same review. Second, we aimed to provide an updated and comprehensive synthesis by including studies with different designs. Three types of design were included: experimental studies (group-comparison studies with control groups and randomisation), quasi-experimental studies (group-comparison studies where there was no randomisation), and studies with a repeated measures design.

1.4. The moderators of effects of self-assessment and peer-assessment interventions

Despite a generally positive relationship between SA/PA interventions and academic performance, the magnitude of the effects varies across studies (Brown & Harris, 2013; Double et al., 2020; Li et al., 2020), with some studies even reporting a non-significant impact (e.g., Covill, 2010; Sadler & Good, 2006). A likely explanation is that the variation in the design and implementation of SA/PA interventions as well as the contexts where the interventions take place influence the effect (Panadero et al., 2017). Thus, efforts to identify the moderators are necessary as they will inform the design of future interventions by revealing the conditions under which interventions are most effective. Based on theory, prior narrative reviews and empirical evidence, we included the characteristics of the implementation of SA/PA, the characteristics of the sample, and the methodology of the included studies as potential moderators.

1.4.1. Moderators related to characteristics of the implementation of self-assessment/peer-assessment

Two groups of moderators related to the implementation of SA and PA were explored. First, students may create and provide quantitative (e.g., scores) or qualitative evaluations (e.g., comments) in SA and PA. Feedback research suggests that, when compare against one another, quantitative scores tend to inhibit further learning because they are seen as summative, while qualitative comments are deemed to be more supportive of learning (Lipnevich et al., 2016). We examined whether this factor (*quantitative or qualitative evaluation*) moderated the intervention effects.

Second, the impact of supporting measures on the successful implementation of SA and PA has been widely studied. For instance, previous research showed that online technology could reduce the logistical burden associated with PA implementation (e.g., Tannacito & Tuzi, 2002; Wen & Tsai, 2008). In addition, the use of online technology may trigger different cognitive and interpersonal demands for SA and PA processes which, in turn, may affect the intervention effects. We, therefore, included the use of online technology as a potential moderator. Student training is regarded as useful for better implementation of and greater learning gains from SA and PA (Brown & Harris, 2013; Li et al., 2016). Sebba et al. (2008) argued that teacher training facilitated the implementation of SA and PA. Miller and Geraci (2011) found that teachers' feedback on students' SA improved SA accuracy for undergraduate students. Teacher feedback is also crucial to improve the quality of the PA or peer feedback (Han & Xu, 2020; Wanner & Palmer, 2018) which, in turn, may influence the effect of PA on students' academic performance. Using instruments (e.g., rubrics, scripts, models,

checklists) is theoretically beneficial for improving the effects of SA and PA because they help improve the reliability, validity, and perceived fairness of assessments conducted by students (Panadero et al., 2013). The empirical evidence was, however, somewhat mixed. In Brown and Harris' (2013) review, SA using rubrics resulted in very large effect sizes and very small (and even negative) effect sizes. Thus, *the use of online technology, student training, teacher training, teacher feedback, and the use of instruments* were included in this review as potential moderators.

1.4.2. Moderators related to characteristics of the sample

The characteristics of the sample influence the effects of the educational interventions (Nelson et al., 2003; Pistone et al., 2019). It has been found that the accuracy of SA and PA can increase with age and schooling experience (Brown et al., 2015; Falchikov & Boud, 1989), probably due to students' increased academic abilities and self-regulation competencies. However, a recent study exploring in detail these effects identified different uses of SA strategies and criteria depending on the year level (Panadero et al., 2020). Therefore, as the direction of results is not yet clear enough, we decided to explore which group of students can gain more from SA and PA interventions. Some studies (e.g., Yan, 2018) reported differences in SA practices between male and female students. Thus, participants' *gender, mean age, educational level* (primary, secondary, and higher education), and *year level* were recorded in order to determine whether these variables moderated the intervention effects.

1.4.3. Moderators related to research method characteristics of studies

The methodological characteristics of studies have been shown to affect the effects of educational interventions (McMillan et al., 2013). Likewise, a review not considering methodological factors may result in misleading conclusions (Cooper, 2010). However, recent meta-analyses (e.g., Double et al., 2020; Li et al., 2020; Youde, 2019) on SA and PA attached very limited attention to the moderating impact of the methodological characteristics of studies. Thus, in this meta-analysis, we examined if the effect sizes from intervention studies were varied by the methodological characteristics of the studies (i.e., *type of design, type of sampling, source of outcome measures, report and control for confounders, type and quality of the instrument, response rate,* and *attrition rate*).

1.5. Aim and research questions of the current review

This meta-analytic review aimed to determine the effectiveness of SA and PA interventions in enhancing students' academic performance and identify the factors that moderate their effectiveness. This meta-analysis extends previous reviews by comparing the effect of SA and PA interventions, and provides comprehensive coverage of available evidence across types of SA/PA and education levels. Studies with either experimental/quasi-experimental design or repeated measures design were included in this review. In particular, this review sought to address three major research questions (RQ):

RQ1 What is the overall effect of SA, PA, and SA + PA (mixed) interventions on students' academic performance?
RQ2 What is the difference between the effect of SA and PA interventions on students' academic performance?
RQ3 How do the intervention effects differ by characteristics of the implementation of SA/PA, the sample, or methodological characteristics?

2. Method

2.1. Search strategies

A series of electronic searches were conducted in two databases, namely ERIC and PsycINFO. These two databases were chosen because they were most likely to contain relevant reports in the field of education. The last search was conducted on August 6, 2020, without restriction on the date of report dissemination. The subject term "effect" together with its alternative terms (i.e., *"effect OR impact OR influence OR result OR outcome OR consequence OR contribution"*) was paired separately with individual self-assessment and peer-assessment alternative terms, including: *"self-assessment", "self-evaluation", "self-monitoring", "self-reflection", "self-rating", "self-grading", "self-review", and "self-feedback"; "peer-assessment", "peer-evaluation", "peer-monitoring", "peer-feedback", "peer-review", "peer-rating", "peer-grading", and "peer-reflection"*. After the initial search, a second search was performed using the same database and all subject terms simultaneously. These searches identified 2,000 potentially relevant journal articles and dissertations for consideration.

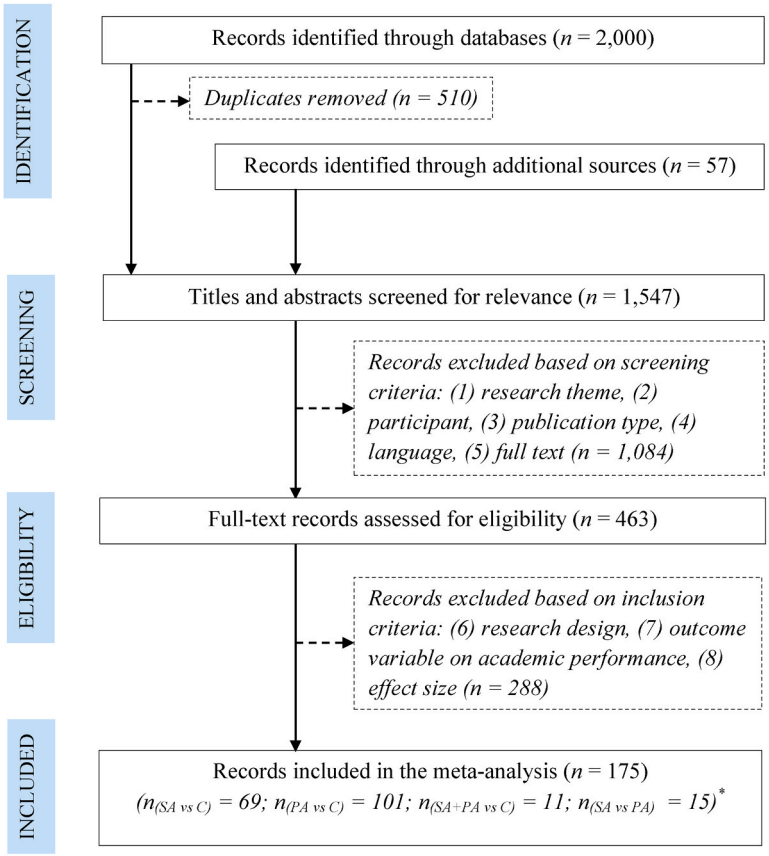
An additional strategy was employed to identify potentially relevant studies that may not have been identified with prior searches in the reference databases. The reviewed studies from eight recent meta-analyses were examined in order to identify potentially relevant studies for the current meta-analysis. Four of them were on SA (Brown & Harris, 2013; Panadero et al., 2017; Papanthymou & Darra, 2019; Youde, 2019), and the other four were on PA (Double et al., 2020; Huisman et al., 2019; Li et al., 2020; Zheng et al., 2020). In total, 330 primary studies were reviewed in these eight meta-analyses. After removing the duplicates, 57 additional studies were added to the existing list for further screening.

2.2. Selection of studies for review

After pooling together and removing duplicate listings from different search strategies, 1,547 separate listings were further screened and filtered based on eight eligibility criteria. These included (1) research theme, (2) participants, (3) publication type, (4)

language, (5) full text, (6) research design, (7) outcome variable, and (8) effect size. Empirical studies examining the effect of SA or PA intervention on student learning (Criterion 1), focusing on students from kindergarten to higher education (excluding students with special educational needs or adult learners attending professional development programs) (Criterion 2), published in peer-reviewed journals or dissertations (Criterion 3), written in English (Criterion 4), and with full text accessible (Criterion 5) were included. To evaluate the effectiveness of a SA or PA intervention, the research design of an included study had to either include a comparison group or serve as their own controls by adopting a repeated measures design (Criterion 6). To clarify, a comparison group could be either a control group (i.e., no SA or PA intervention) or an experimental group of the other assessment type (e.g., SA vs. PA). Next, echoing the research question concerning academic learning outcome, a study to be selected had to include at least one quantitative measure of student academic performance (Criterion 7). Furthermore, the effect size (i.e., the standardised mean difference of intervention) of a study needed to be reported or calculable with reported statistics (mean, sample size, and standard deviation), or data obtained from authors via email in order to be included (Criterion 8).

As a result, 175 studies were included in the meta-analysis (see Table S1 of the supplementary materials for a summary). Among them, 69 studies provided information for comparing SA and a control group, 101 studies for comparing PA and a control group, 11 studies for comparing SA + PA (mixed) and a control group, and 15 studies for comparing PA and SA. A SA + PA (mixed) intervention refers to incorporating both SA and PA activities simultaneously (e.g., peer feedback and self-editing on a writing task). Considering the variation of control groups in the primary studies, we took extra precautions to enhance the homogeneity of control groups. A control group was defined as a baseline group without any SA/PA intervention in the current review, regardless of its original label in primary studies. For example, in a study examining the impact of using online technology in PA, a face-to-face PA group was treated as a control group. However, since the control group received PA intervention, it was re-coded from a control group in the original study to a PA experimental group in this meta-analysis. In brief, the experimental group was the same as the control group except that the experimental group had SA/PA intervention. Given that some studies reported multiple comparisons (e.g., SA vs. Control and PA vs. Control), the sum of all comparisons was bigger than the total number of included studies.



Note. PRISMA = Preferred Reporting Items for Systematic Reviews and Meta-analyses

Fig. 1. PRISMA flow diagram of the search and screening process.
Note. PRISMA = Preferred Reporting Items for Systematic Reviews and Meta-analyses.
*Given that some studies reported multiple pairwise comparisons, the sum of all comparisons was bigger than the total number of included studies.

To ensure the quality of selection, we randomly selected a sample of 125 studies from the identified records to check for any potential misunderstanding before the full screening process. Two researchers screened and filtered these 125 studies independently based on the inclusion and exclusion criteria. Then, their inclusion decisions were compared, resulting in a 91% inter-rater agreement. All disagreements were discussed to reach a consensus before filtering the rest of the records. A PRISMA flow diagram in Fig. 1 illustrates the identification, screening, and filtering processes.

2.3. Data extraction

After the selection process, all the included studies were read thoroughly and coded using a structured data extraction form. The form was constructed, pilot tested and modified by the authors. It included four sections of information: (1) demographic information of participants; (2) intervention moderators; (3) outcome measures; and (4) study quality. To ensure the quality of data extraction, two researchers independently coded all requested information on a randomly selected sample of 60 studies from the included records, resulting in a 94% inter-rater agreement. All disagreements were discussed and resolved by consensus before proceeding to further coding. In case no information was provided in the primary studies, it was coded as missing data. The codebook was refined accordingly. One coder was responsible for the remained studies with detailed documentation to maximise consistency and accuracy. Any uncertainty was resolved by group discussion. After the initial complete coding, a second-round checking was conducted with a random sample of 30 included studies, resulting in a 97% inter-rater agreement.

Demographic information of participants included country, gender, mean age, educational level, and year level. *Intervention moderators* included (a) whether the qualitative evaluation was provided by assessors or received by assessees (e.g., oral or written comments); (b) whether the quantitative evaluation was provided by assessors or received by assessees (e.g., rating/grade/score); (c) whether online technology was employed during SA and PA; (d) whether student assessors received training; (e) whether teachers who taught the participants received training on SA or PA instruction for the study; (f) whether students received any teacher feedback on SA or PA process or outcome; and (g) whether students used any assessment instrument during SA or PA (e.g., scoring rubric, evaluation form, checklist, prompt). *Outcome measures* included both cognitive knowledge (e.g., final test score) and skills (e.g., oral presentation performance). *Study quality* was assessed through a modified version of the Effective Public Health Practice Project (EPHPP) instrument (Thomas et al., 2004). This original instrument covers the following dimensions of quality: selection bias, study design, confounders, blinding, data collection methods, and withdrawals and dropouts. We excluded the "blinding" dimension for this meta-analysis because blinding participants or researchers to the conditions is challenging in educational studies (Noetel et al., 2021), and studies reporting this information are rare. More information about how each item of this instrument was coded and about how the final quality score was computed can be found in the supplementary materials.

2.4. Statistical analyses

Cohen's *d* was calculated for those studies that compared interventions *(SA vs control, PA vs control, SA + PA vs control, SA vs PA)*. For other studies with a repeated measures design, we calculated the Cohen's *d* using the formula of Morris and DeShon (2002, p. 117), which is comparable to the Cohen's *d* calculated from studies comparing interventions. A drawback of the formula of Morris and DeShon (2002) is that we need to know the correlation between the performance scores on both time points, but this correlation was never reported in the associated studies. Therefore, we imputed a correlation of 0.50. Furthermore, Cohen's *d* effect size might be overestimated when the sample size is small. To correct for this bias, Cohen's *d* was converted into Hedges' *g* (Hedges, 1981). A positive effect size indicates that PA or SA works better than the control condition. When comparing SA and PA, a positive effect size indicates that PA works better than SA.

Studies often included several effect sizes, either because they measured several academic outcomes, or because several interventions were implemented (i.e., SA, PA, and control), generating dependency among them. This dependency was taken into account by applying a three-level model (Cheung, 2014; Van den Noortgate et al., 2013). More information about this model can be found in the supplementary materials. Moderator variables were included to explain the variance observed at each level. Four separate analyses were done for each of the four comparisons: 1) SA – control, 2) PA – control, 3) SA + PA (mixed) – control, and 4) SA – PA. If there were less than 10 effect sizes for a given moderator analysis, a univariate random-effects meta-analysis was fitted.

To test the robustness of our findings, we checked for the existence of outliers. For each comparison, effect sizes that were beyond 2 standard deviations above or below the mean were considered outliers, and analyses were repeated without these effects. Finally, we also checked for the existence of publication bias through the visual analysis of funnel plots (Light & Pillemer, 1984), and through the three-level Egger regression test (Egger et al., 1997; Fernández-Castilla et al., 2020). If evidence of publication bias was detected, the selection model of Vevea and Woods (2005) was applied in order to achieve an overall estimate adjusted for the presence of publication bias. More information about this selection model can be found in the supplementary materials. All the analyses were performed in R, using *metafor* (Viechtbauer, 2010) and *weightr* (Coburn & Vevea, 2019).

3. Results

In this section, we present the results for four major comparisons: 1) SA – control, 2) PA – control, 3) SA + PA (mixed) – control, and 4) SA – PA. For each comparison, we first present the overall effect, followed by moderator analysis results. As outliers were detected in most of the comparisons and some of these outliers were abnormally large, we present the results without outliers in this section because they are more robust.

3.1. SA – control comparison

A total number of 227 effect sizes within 69 studies reported a comparison between a control group and a group that implemented an SA intervention, or a pre-post comparison of SA. In this group of studies, 12 outliers were detected (Hedges' g were above 4.60 or below −2.89). The overall effect of SA intervention (compared to control) was 0.584, that was statistically different from zero (SE = 0.096, Z = 6.105, p-value < .001, k = 215 in 66 studies). This overall effect significantly varied across studies (between-studies variance = 0.516; LRT: X^2 = 116.15, p-value < .001) and within studies (within-study variance = 0.080; LRT: X^2 = 100.45, p-value < .001). The results of the meta-regression (without the outliers) are presented in Table S2.1 and Table S2.2 of the supplementary materials. When the type of design was repeated measures, the overall effect was significantly higher (g = 1.084) than when the design was experimental – group comparison (g = 0.363, p-value of the difference 0.014) and when the design was quasi-experimental – group comparison (g = 0.476, p-value of the difference 0.007). Also, when the quality of the studies was moderate, the overall effect was significantly smaller (g = 0.246) than when the quality of the studies was weak (g = 0.699, p-value of the difference <0.05). However, the overall effect of high-quality studies was similar to those of low-quality (g = 0.865, k = 10). Therefore, we do not consider that the quality of studies alone influences the overall effect. None of the other moderators was significant.

3.2. PA – control comparison

Regarding the number of effects that compared PA with a control group or that reported a pre-post comparison of PA, 349 effect sizes were coded within 101 studies. For this subset of studies, 11 outliers were detected (Hedges' g were above 4.45 or below −2.74). The overall effect of PA (compared to control) was 0.613, that was statistically different from zero (SE = 0.076, Z = 8.025, p-value < .001, k = 338 in 99 studies). This overall effect did vary across studies (between-studies variance = 0.498; X^2 = 198.35, p-value <. 001) and within-studies (within-study variance = 0.086; LRT: X^2 = 127.56, p-value < .001). In those studies where technology was used, the overall effect was statistically larger (g = 0.871) than when online technology was not used (g = 0.444, p-value of the difference <0.01). The type of design used across studies influenced the overall effect size: in experimental-group comparison studies, the overall effect was smaller (g = 0.290) than that in repeated measures studies (g = 0.877, p-value of the difference <0.05). No significant effect was found for the other moderators.

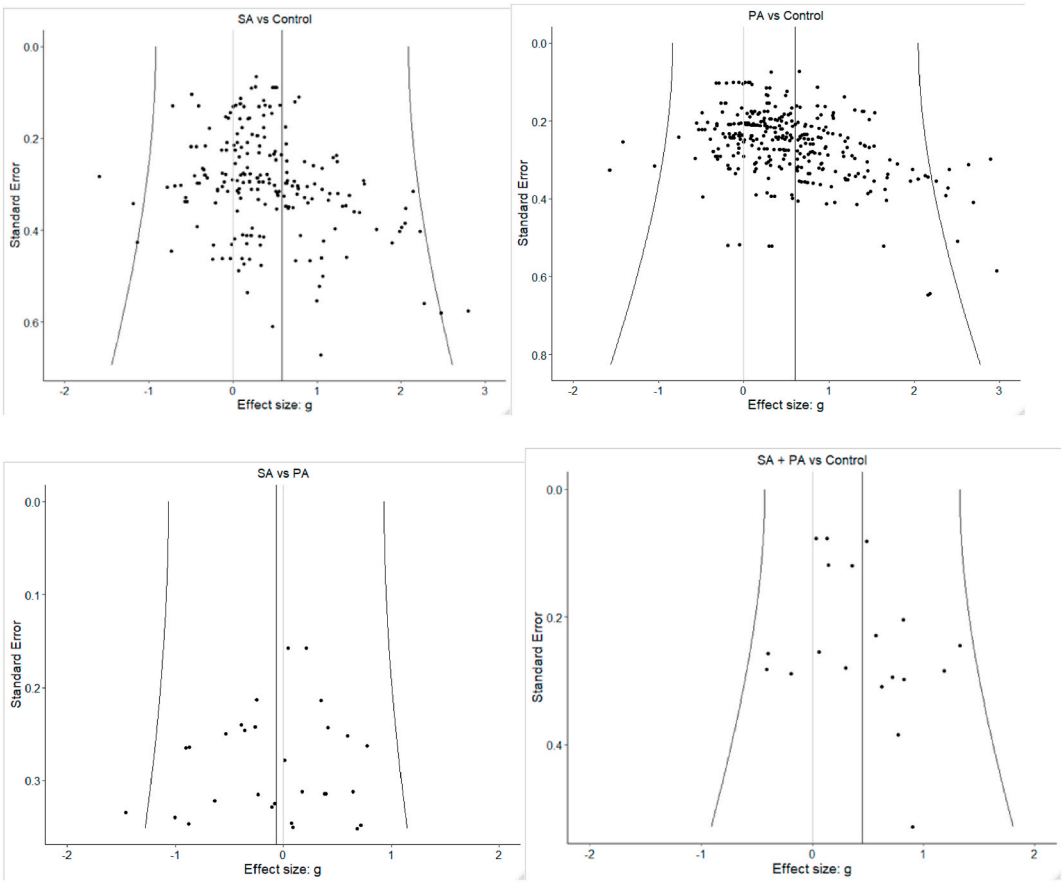


Fig. 2. Funnel plot for each pairwise comparison.

3.3. SA + PA – control comparison

A total of 19 effect sizes within 11 studies reported a comparison between a control group and a SA + PA group, or a pre-post comparison of SA + PA. In this case, no outliers were detected. The overall effect of the combination of PA and SA (compared to a control) was 0.448, that was statistically different from zero (SE = 0.146, Z = 3.074, p-value < .05; k = 19 within 11 studies). This overall effect did not significantly differ across studies (between-studies variance = 0.159; LRT: X^2 = 3.433, p-value = .064) but did differ within studies (within-study variance = 0.042; LRT: X^2 = 12.219, p-value < .01). There was a positive effect of the mean age of the participants (B = 0.338, SE = 0.159, Z = 2.128, p < .05, k = 6), meaning that the higher the mean age, the higher the overall effect (or the larger the differences between the control group and the PA + SA group). No significant effect was found for the other moderators.

3.4. SA – PA comparison

A total number of 31 effect sizes within 15 studies reported a direct comparison between SA and PA. In this case, two outliers were detected (Hedges' g were below −2.00). An overall effect of −0.065 was observed, which was not statistically different from zero (SE = 0.144, Z = −0.453, p-value = .65; k = 29 within 13 studies). There was significant variation between studies (between-studies variance = 0.208; LRT: X^2 = 5.897, p-value < .05) but not within studies (within-study variance = 0.051; LRT: X^2 = 1.860, p-value = .173). None of the moderators was significant.

3.5. Publication bias

The potential publication bias was checked with the funnel plots and the three-level Egger regression test. The funnel plots of each pairwise comparison can be found in Fig. 2. Asymmetries in the funnel plots can be observed for the comparison "SA-control" and for the comparison "PA-control", where less precise studies (i.e., studies with smaller sample sizes) show larger effects favouring SA and PA, respectively. This asymmetry was confirmed by the three-level Egger regression test (SA-control: B = 2.94, p-value < .001; PA-control: B = 6.27, p-value < .0001). The selection model of Vevea and Woods (2005) indicated that the adjusted pooled effect of SA (compared to control) assuming moderate publication bias was 0.343 and the adjusted pooled effect of PA (compared to control) was 0.406. Therefore, it is possible that the overall effects observed for these comparisons are somewhat inflated. The funnel plots for the comparisons "SA + PA vs Control" and "SA vs PA" were more symmetrical, and the three-level Egger regression test did not indicate the presence of bias.

4. Discussion

To our knowledge, this study represents the most comprehensive meta-analysis to date about SA/PA interventions and their effects on academic performance. Our meta-analysis included 626 effect sizes from 175 independent studies involving a total of 19,383 participants. The results presented here were based on 601 effect sizes from 169 studies involving 19,003 participants after removing outliers. We discuss our findings relevant to three research questions: RQ1-What is the overall effect of SA, PA, and the SA + PA (mixed) interventions on students' academic achievement? RQ2-What is the difference between the effect of SA and PA interventions? RQ3-What are possible moderators of the intervention effects?

4.1. Overall effect of self-assessment, peer-assessment, and the mixed interventions

Results indicated that SA intervention (g = 0.585), PA intervention (g = 0.606), and SA + PA (mixed) intervention (g = 0.448) led to significantly better academic performance. These effect sizes could be regarded as medium according to Cohen's (1969) general benchmarks. However, Cohen's categories of small, medium and large effect size might not be appropriate for interpreting the effects of intervention studies in education (Lipsey et al., 2012), and researchers have recommended different benchmarks. For example, Hattie (2008) described an effect size of 0.40 as a benchmark for meaningful interventions and an effect size of 0.60 as large. Kraft (2020) proposed that less than 0.05 is small, 0.05 to less than 0.20 is medium, and 0.20 or greater is large in terms of the effect of educational interventions. Therefore, the overall mean effect size revealed in this meta-analysis can be interpreted as meaningful and large within the context of educational interventions. These results testify to the theoretical arguments for the usefulness of SA and PA (Andrade, 2019; Dochy et al., 1999; Harris & Brown, 2018; Panadero et al., 2016, 2017; Topping, 1998, 2003) and suggest that SA and PA can effectively enhance students' academic performance in a wide range of education levels and age groups.

The overall mean effect sizes revealed in this meta-analysis are larger than those reported in previous meta-analyses, where the effect sizes range from 0.40 to 0.46 for SA interventions (Brown & Harris, 2013; Youde, 2019), and 0.26 to 0.33 for PA interventions (Double et al., 2020; Li et al., 2020). This result should be interpreted with caution. A viable explanation is that the current meta-analysis included studies with experimental/quasi-experimental design or repeated measures design, while most previous meta-analyses covered only experimental/quasi-experimental studies. For both SA and PA interventions, experimental/quasi-experimental studies had smaller effect sizes than those in repeated measures studies. Another possible explanation is the existence of publication bias. Previous meta-analyses either reported no presence of publication bias or did not report it. However, as reported in Youde's (2019) review, unpublished studies, which represented more than half of the studies included,

demonstrated a lower mean effect compared with published studies. Following the recommendations by Banks et al. (2012), we included both published (peer-reviewed journal articles) and unpublished studies (dissertations and theses) in this meta-analysis. However, the presence of publication bias suggests that the current meta-analysis might present a relatively optimistic picture of the effect of SA/PA interventions. The adjusted pooled effect of SA assuming moderate publication bias was 0.343, and the adjusted pooled effect of PA was 0.434, which are closer to the magnitude of effect sizes reported in previous meta-analyses.

SA and PA are often implemented together in practice and are supposed to inform and enhance each other (Dochy et al., 1999; Sanchez et al., 2017). However, this argument seems more theory-grounded than empirically supported in this meta-analysis. Although the combination of SA + PA intervention resulted in significantly better academic performance ($g = 0.448$) compared with control groups, the effect size of the combined intervention was smaller than that of individual SA or PA intervention, indicating combining SA and PA in the same intervention did not increase performance as much as doing SA or PA individually. In fact, there is one study (Birjandi & Tamjid, 2012) that directly compared the effect of the SA + PA intervention against SA and PA alone. The results showed that PA alone was superior to PA + SA ($g = 1$) and that SA alone was slightly better than SA + PA ($g = 0.19$). It is plausible that combining SA and PA together increased the complexity of the intervention, which is more cognitively demanding for participants. The increased cognitive load imposed by complex learning tasks may seriously impair learning (van Merriënboer & Sluijsmans, 2009). van Zundert et al. (2012) found that combining instruction of peer assessment with complex tasks might cause a high cognitive load and, therefore, hampers student learning. Similarly, researchers (e.g., Lipnevich et al., 2014; Lipnevich et al., 2022) found that using rubrics and exemplars one-on-one worked better than using both of them at the same time. However, as there is no direct evidence regarding the relationship between cognitive load and the complexity of the SA + PA intervention, this explanation remains speculative.

4.2. Difference between the effect of self-assessment and peer-assessment interventions

When SA and PA interventions were conducted within the same study, their effects were not significantly different. This is consistent with prior meta-analyses. For instance, Li et al. (2020) identified 20 effect sizes directly comparing SA and PA interventions, and Double et al. (2020) covered 10 studies with direct comparisons between SA and PA groups. Both reviews found no significant difference between the effect of SA and PA interventions. This finding suggests that when implemented appropriately, both SA and PA can enhance students' academic performance in a similar magnitude (Panadero et al., 2016). The choice between SA and PA should be based on considerations such as the target learning competencies and skills, and students' levels of maturity or previous experience with SA and PA. Although it may involve others, SA is mainly an inwards: it requires a high level of self-awareness and self-regulation because during the process students need to set their learning targets, gather feedback from different sources and reflect on performance (Butler & Winne, 1995; Panadero, Broadbent, et al., 2019). In contrast, PA requires students to evaluate each other's work and involves interpersonal processes (van Gennip et al., 2009). Students can learn in the roles of assessor or assessee in PA because providing and receiving feedback are precious learning opportunities (Panadero et al., 2018).

Importantly, SA and PA are associated with different challenges. SA has been criticised as an inaccurate indicator of student performance (Brown et al., 2015). Students tend to either overrate themselves due to overconfidence in newly learned skills or a self-serving bias or underestimate themselves because of negative illusions (Dunning et al., 2004). PA might be less subjective to a self-serving bias or negative illusion, but it has its own problems, such as overestimation due to friendship marking or lack of differentiation due to collusive marking (Pond et al., 1995). Such flaws in SA and PA hinder their function of learning support. While students might not be accurate in their SA/PA, it is considered that part of the instructional benefit of involving them in SA/PA is precisely to turn them more accurate (Panadero et al., 2016). Veridicality is still an essential and worthwhile goal to pursue in SA/PA because realistic judgements help students focus on their learning needs (Yan, 2022).

4.3. Moderators of self-assessment and peer-assessment interventions

In this meta-analysis, we explored three sets of moderators. Firstly, we did not identify significant intervention moderators related to the implementation of SA. For PA interventions, only the use of online technology increased the overall effect. Previous research (e.g., Tannacito & Tuzi, 2002; Wen & Tsai, 2008) argued that online technology could facilitate PA by reducing the logistical burden. Unlike SA, which is mainly an introspective process, PA requires more interpersonal interactions that can be well supported by online technology. This may explain why using online technology is a significant moderator for the effect of PA, but not SA interventions. This finding (no significant moderator for SA interventions and only one for PA interventions) seems to suggest that the design or the implementation of SA/PA interventions does not affect their effects on learning performance. However, we are inclined to make a cautious interpretation of the results, similarly to what Panadero et al. (2017) claimed for their review on the effect of SA on self-regulated learning. That is, the moderators explored in this review showing non-significant impact does not necessarily mean all SA/PA interventions are equally effective. As formative assessment practices are influenced by both personal and contextual factors (Yan et al., 2021), it is less likely to have a one-fit-to-all solution. While the current review attempted to provide an overall picture of the effect of SA/PA across contexts, future reviews may explore in more depth and with more homogenous studies, with the aim of identifying the features of effective SA/PA interventions.

Secondly, some sample characteristics were found to moderate the effect of the SA + PA interventions: participants with higher mean age had more learning gains. Considering the possible relationship between cognitive load and the complexity of the intervention, as discussed earlier, this finding is not a surprise. The extra cognitive load resulting from the increased complexity of the intervention might have a more negative impact on younger participants than their older counterparts.

Thirdly, the research design matters for both SA and PA intervention studies. Studies with a repeated measures design generated

much larger effect sizes than studies that used an experimental and quasi-experimental design. A similar difference favouring repeated measures design has been found in previous reviews on different topics (e.g., Kroesbergen & van Luit, 2003; Roehling et al., 2013). There are several viable reasons. It is possible that, in studies with repeated measures design, the same test or parallel tests were used so that the practising effect resulted in inflated performance. Another speculative explanation is that the baseline test in repeated measures designs usually involved no intervention, while the control group in some experimental/quasi-experimental designs might have received some form of intervention different from SA/PA. Hence, the effect of SA/PA in experimental/quasi-experimental designs might be underestimated. This finding reminds us about the purposes of using placebo or no-treatment control groups (i.e., groups without any intervention) and active control groups (i.e., groups receiving intervention other than SA/PA). Similar to Ovosi et al.'s (2017) argument, the former tells about the absolute efficacy of SA/PA, while the latter serves to differentiate the relative efficacy of SA/PA compared with other interventions. In addition, some studies with experimental and, most especially, quasi-experimental designs did not control confounders in order to assure comparability among groups. The results, therefore, might be contaminated.

In sum, many of the hypothesised moderators did not significantly moderate the intervention effects. Although theoretical arguments and individual empirical studies supported the important role of the hypothesised moderators, little evidence was found in this meta-analysis regarding their moderating effect. This result echoes recent meta-analysis reviews that failed to identify many significant moderators. For instance, in Youde's (2019) review on SA, educational level was found to be the only significant moderator - i.e., stronger effect for middle school learners than high school learners. Double et al. (2020) found no significant moderator for PA interventions. Li et al. (2020) identified only one significant moderator for PA interventions. That is, students receiving rater training had larger learning gains than those who did not receive such training. However, caution is needed in the interpretations of this finding. Firstly, the small number of studies in some subgroup analyses (e.g., SA + PA – control comparison and SA – PA comparison) may reduce the statistical power to detect significant moderators. Secondly, the collinearity of moderators (Murano et al., 2020), i.e., multiple moderators nested within particular intervention programs, may leave intervention programs as the primary cause of outcomes but not give valid results about any individual moderator.

4.4. Implications for research and practice

Researchers argue that SA and PA are helpful for students to identify the performance gap between their current and desired levels so that they can better direct their further learning (Andrade, 2010; Boud et al., 2013; Topping, 1998). SA and PA are also believed to facilitate both self-regulation and co-regulation and, therefore, bring about better learning outcomes (Panadero, Broadbent, et al., 2019; Yan et al., 2020). This meta-analysis corroborates these theoretical arguments and empirical investigations about the positive impact of SA and PA on students' academic performance across contexts. Importantly, SA and PA are less vulnerable to the typical practical constraints associated with teacher-directed formative assessment (e.g., big class sizes) (Yan & Brown, 2021) since the teacher is no longer the sole source of feedback and every individual student becomes learning resources for themselves and one another (Black & Wiliam, 2009). Hence, we argue that SA and PA should be used more frequently in classrooms and, if possible, be embedded in the official curriculum so that students can develop skills in SA and PA in a systematic and cumulative approach.

Appropriate design and implementation are crucial because SA or PA does not always enhance academic performance. Although the overall effect of SA and PA was positive, negative effects were observed in 22.79% of SA and 20.35% of PA interventions among the studies included in this meta-analysis. As revealed in a recent narrative review (Yan et al., 2021), the implementation of formative assessment, including SA and PA, is complex and influenced by a wide range of personal and contextual factors, as well as the interactions among those factors. Unfortunately, not many significant moderators were identified in this meta-analysis except that using online technology appeared to be a good strategy to increase the effect of PA interventions. Teachers are suggested to consider both personal factors (e.g., whether teachers and students are ready?) and contextual factors (e.g., are necessary support available?) when designing and carrying out SA and PA in classrooms.

The results of this meta-analysis suggest that we need to consider that conducting SA or PA together results in a slightly lower improvement of academic performance. Compared with traditional teacher-directed assessment, SA and PA processes are more complicated and cognitively demanding to students. Although theoretically, SA and PA are supposed to inform and enhance each other, in practice, combining them may reduce the positive impact on academic performance, probably due to the excessive cognitive load. Thus, students' abilities or readiness should be an important factor to consider where a combined SA + PA intervention is preferred. Since SA and PA can support the development of a wide range of capacities, each intervention may focus on a particular subset of capacities and provide appropriate supporting measures (such as teacher feedback and online technology) to reduce the cognitive load caused by elements that are out of the foci.

4.5. Limitations and future research

There are several limitations to note. First, some subgroup analyses may have been underpowered due to the limited number of studies and effect sizes. For instance, only 31 effect sizes, including two outliers, within 15 studies directly compared the effects of SA and PA. As more studies on this topic accumulate over time, future meta-analyses may have more statistical power to detect potentially meaningful moderators that have not been identified in this analysis.

Second, we have used traditional pairwise meta-analytic techniques to carry out the present meta-analysis, but more advanced techniques, such as network meta-analysis (Salanti, 2012), could be considered in the future. However, the heterogeneity of the study characteristics might be a challenge for complying with its assumptions (i.e., transitivity and consistency).

Third, a long-lasting criticism of meta-analyses is the impact of publication bias. For both the SA-control and PA-control

comparison in this meta-analysis, studies with smaller sample sizes showed larger effects, indicating that the overall effects observed might be inflated by less precise studies. To minimise publication bias, future meta-analyses could consider including more types of unpublished studies in addition to dissertations and theses, such as book chapters and conference proceedings. Researchers should also pay more attention to the process of reporting research data in order to ensure the rigour of their studies.

Fourth, this meta-analysis used only academic performance as the outcome variable. This decision was taken mainly because the academic performance was the prevailing interest of SA and PA intervention studies. However, there is also emerging evidence regarding the effect of SA and PA on promoting learning strategies and social and emotional learning outcomes, such as self-regulation and self-efficacy (Nieminen et al., 2019; Panadero et al., 2017), intrinsic value and motivation (Yan et al., 2020), learner autonomy (Shen et al., 2020), positive thinking (Wang et al., 2017), and pre-service teachers' professional vision of classroom management (Weber et al., 2018). Future meta-analyses could depict a more comprehensive picture of the potential of SA and PA by covering both academic and non-academic outcomes.

Lastly, it is worthwhile to mention that many studies included in this meta-analysis were classified as low-quality studies, which directly affects the reliability of their findings. Future applied researchers should use better designs (i.e., experimental designs) to test the effectiveness of SA/PA interventions, improve the sampling method (i.e., random sampling), and also improve the reporting of the control of confounding variables.

5. Conclusion

The current meta-analysis provided a comprehensive and updated synthesis of the effect of SA and/or PA interventions on academic performance based on data from 626 effect sizes from 175 independent studies involving a total of 19,383 participants. Our results corroborated that SA and PA interventions can enhance students' academic performance across contexts. When SA and PA interventions were examined in the same study, their effects did not significantly differ. SA + PA (mixed) intervention also improved students' academic performance but the effect was slightly smaller than SA or PA alone. Three moderators influenced our results: (a) online technology increased the effect of PA on performance, (b) participants with higher mean age had more learning gains from SA + PA (mixed) intervention, and (c) studies with repeated measures designs generated larger effect sizes than studies with experimental/quasi-experimental design for both SA and PA. The findings of the current meta-analysis lend credit to the use of the SA and PA interventions in classrooms and could inform the design and implementation of SA and PA interventions. Future meta-analytic work may consider synthesising the effects of SA and PA on promoting learning strategies and social and emotional learning outcomes.

CRediT authorship contribution statement

Zi Yan: Conceptualization, Data curation, Writing – original draft, Writing – review & editing, Funding acquisition. **Hongling Lao:** Conceptualization, Data curation, Writing – original draft, Writing – review & editing. **Ernesto Panadero:** Conceptualization, Data curation, Writing – original draft, Writing – review & editing. **Belen Fernández-Castilla:** Data curation, Formal analysis, Writing – original draft, Writing – review & editing. **Lan Yang:** Data curation, Writing – review & editing. **Min Yang:** Data curation, Writing – review & editing.

Declarations of competing interest

None.

Data availability

Data will be made available on request.

Acknowledgements

The work described in this paper was supported by a General Research Fund from the Research Grants Council of the Hong Kong SAR, China (Project No. EDUHK 18600019).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.edurev.2022.100484.

References

Allal, L. (2016). The co-regulation of student learning in an assessment for learning culture. In L. Allal, & D. Laveault (Eds.), *Assessment for learning: Meeting the challenge of implementation* (pp. 259–273). Springer.

Andrade, H. L. (2010). Students as the definitive source of formative assessment: Academic self-assessment and the self-regulation of learning. In H. Andrade, & G. Cizek (Eds.), *Handbook of formative assessment* (pp. 90–105). Routledge.
Andrade, H. L. (2019). A critical review of research on student self-assessment. *Frontiers in Education, 4*. https://doi.org/10.3389/feduc.2019.00087
Andrade, H. L., & Heritage, M. (2018). *Using formative assessment to enhance learning*. Routledge. achievement, and academic self-regulation.
Baas, D., Castelijns, J., Vermeulen, M., Martens, R., & Segers, M. (2015). The relation between Assessment for Learning and elementary students' cognitive and metacognitive strategy use. *British Journal of Educational Psychology, 85*(1), 33–46.
Banks, G. C., Kepes, S., & Banks, K. P. (2012). Publication bias: The antagonist of meta-analytic reviews and effective policymaking. *Educational Evaluation and Policy Analysis, 34*, 259–277. https://doi.org/10.3102/0162373712446144
Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*, 7–74.
Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5–31.
Boud, D. (1991). *Implementing student self-assessment*. Higher Education Research and Development Society of Australasia.
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.
Brown, G. T. L., Andrade, H. L., & Chen, F. (2015). Accuracy in student self-assessment: Directions and cautions for research. *Assessment in Education: Principles, Policy & Practice, 22*(4), 444–457. https://doi.org/10.1080/0969594x.2014.996523
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). Sage.
Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*(3), 245–281.
Cahyono, B. Y., & Amrina, R. (2016). Peer feedback, self-correction, and writing proficiency of indonesian EFL students. *Arab World English Journal, 7*(1), 178–193. https://doi.org/10.24093/awej/vol7no1.12
Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education, 43*(8), 1315–1325.
Cheng, K. -H., Liang, J. -C., & Tsai, C. -C. (2015). Examining the role of feedback messages in undergraduate students' writing performance during an online peer assessment activity. *The Internet and Higher Education, 25*, 78–84.
Cheung, M. W. L. (2014). Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach. *Psychological Methods, 19*(2), 211.
Coburn, K. M., & Vevea, J. K. (2019). *weightr: Estimating weight-function models for publication bias. R package version 2.0.2* https://CRAN.R-project.org/package=weightr.
Cohen, J. (1969). *Statistical power analysis for the behavioral sciences* (1st ed.). Academic Press.
Cooper, H. (2010). *Research synthesis and meta-analysis*. Sage.
Covill, A. E. (2010). Comparing peer review and self-review as ways to improve college students' writing. *Journal of Literacy Research, 42*, 199–226. https://doi.org/10.1080/10862961003796207
Diab, N. M. (2011). Assessing the relationship between different types of student feedback and the quality of revised writing. *Assessing Writing, 16*(4), 274–292. https://doi.org/10.1016/j.asw.2011.08.001
Dochy, F. J., & McDowell, L. (1997). Introduction: Assessment as a tool for learning. *Studies In Educational Evaluation, 23*(4), 279–298.
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
Double, K. S., McGrane, J. A., & Hopfenbeck, T. N. (2020). The impact of peer assessment on academic performance: A meta-analysis of control group studies. *Educational Psychology Review, 32*(2), 481–509. https://doi.org/10.1007/s10648-019-09510-3
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.
Egger, M., Davey-Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*, 629–634.
Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research, 59*(4), 395–430.
Fernández-Castilla, B., Declercq, L., Jamshidi, L., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2021). Detecting selection bias in meta-analyses with multiple outcomes: A simulation study. *The Journal of Experimental Education, 89*(1), 125–144.
van Gennip, N. A., Segers, M. S., & Tillema, H. H. (2009). Peer assessment for learning from a social perspective: The influence of interpersonal variables and structural features. *Educational Research Review, 4*(1), 41–54.
Hadwin, A. F., & Oshige, M. (2011). Self-regulation, co-regulation, and socially-shared regulation: Exploring perspectives of social in self-regulated learning theory. *Teachers College Record, 113*(2), 240–264.
Han, Y., & Xu, Y. (2020). The development of student feedback literacy: The influences of teacher feedback on peer feedback. *Assessment & Evaluation in Higher Education, 45*(5), 680–696. https://doi.org/10.1080/02602938.2019.1689545
Harris, L. R., & Brown, G. T. L. (2018). *Using self-assessment to improve student learning*. Routledge.
Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.
Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107–128.
Huisman, B., Saab, N., van den Broek, P., & van Driel, J. (2019). The impact of formative peer feedback on higher education students' academic writing: A meta-analysis. *Assessment & Evaluation in Higher Education, 44*(6), 863–880. https://doi.org/10.1080/02602938.2018.1545896
Kraft, M. (2020). Interpreting effect sizes of education interventions. *Educational Researcher, 49*(4), 241–253. https://doi.org/10.3102/0013189X20912798
Kroesbergen, E. H., & van Luit, J. E. H. (2003). Mathematics interventions for children with special educational needs: A meta-analysis. *Remedial and Special Education, 24*(2), 97–114.
Li, H., Xiong, Y., Zang, X., Kornhaber, M., Lyu, Y., Chung, K., & Suen, H. K. (2016). Peer assessment in a digital age: A meta-analysis comparing peer and teacher ratings. *Assessment & Evaluation in Higher Education, 41*(2), 245–264. https://doi.org/10.1080/02602938.2014.999746
Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Harvard University Press.
Lipnevich, A. A., Berg, D. A. G., & Smith, J. K. (2016). Toward a model of student response to feedback. In G. T. L. Brown, & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 169–185). Routledge.
Lipnevich, A. A., McCallen, L. N., Miles, K. P., & Smith, J. K. (2014). Mind the gap! Students' use of exemplars and detailed rubrics as formative assessment. *Instructional Science, 42*(4), 539–559. https://doi.org/10.1007/s11251-013-9299-9
Lipnevich, A. A., Panadero, E., & Calistro, T. (2022). Unraveling the effects of rubrics and exemplars on student writing performance. *Journal of Experimental Psychology*. https://doi.org/10.1037/xap0000434.
Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., & Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*. National Center for Special Education Research.
Lee, C. -I., Yang, Y. -F., & Mai, S. -Y. (2016). The impact of a scaffolded assessment intervention on students' academic achievement in web-based peer assessment activities. *International Journal of Distance Education Technologies, 14*(4), 41–54.
Li, H., Xiong, Y., Hunter, C. V., Guo, X., & Tywoniw, R. (2020). Does peer assessment promote student learning? A meta-analysis. *Assessment & Evaluation in Higher Education, 45*(2), 193–211. https://doi.org/10.1080/02602938.2019.1620679
McMillan, J. H., Venable, J. C., & Varier, D. (2013). Studies of the effect of formative assessment on student achievement: So much more is needed. *Practical Assessment, Research and Evaluation, 18*(2), 1–15.
van Ginkel, S., Gulikers, J., Biemans, H., & Mulder, M. (2017). The impact of the feedback source on developing oral presentation competence. *Studies in Higher Education, 42*(9), 1671–1685.
van Merriënboer, J. J. G., & Sluijsmans, D. M. A. (2009). Toward a synthesis of cognitive load theory, four-component instructional design, and self-directed learning. *Educational Psychology Review, 21*, 55–66. https://doi.org/10.1007/s10648-008-9092-5

Meusen-Beekman, K. D., Joosten- ten Brinke, D., & Boshuizen, H. P. A. (2015). Developping self-regulation in primary education by means of formative assessment: A theoretical perspective. *Cogent Education., 2*(1), 1–16. https://doi.org/10.1080/2331186x.2015.1071233
Miller, T. M., & Geraci, L. (2011). Training metacognition in the classroom: The influence of incentives and feedback on exam predictions. *Metacognition and Learning, 6*, 303–314. https://doi.org/10.1007/s11409-011-9083-7
Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7*(1), 105.
Murano, D., Sawyer, J. E., & Lipnevich, A. A. (2020). A meta-analytic review of preschool social and emotional learning interventions. *Review of Educational Research, 90*(2), 227–263. https://doi.org/10.3102/0034654320914743
Nelson, J. R., Benner, G. J., & Gonzalez, J. (2003). Learner characteristics that influence the treatment effectiveness of early literacy interventions: A meta-analytic review. *Learning Disabilities Research & Practice, 18*(4), 255–267.
Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning. *A model and seven principles of good feedback practice. Studies in Higher Education, 31*(2), 199–218. https://doi.org/10.1080/03075070600572090
Nieminen, J. H., Asikainen, H., & Rämö, J. (2019). Promoting deep approach to learning and self-efficacy by changing the purpose of self-assessment: A comparison of summative and formative models. *Studies in Higher Education, 46*(7), 1296–1311. https://doi.org/10.1080/03075079.2019.1688282
Noetel, M., Griffith, S., Delaney, O., Sanders, T., Parker, P., del Pozo Cruz, B., & Lonsdale, C. (2021). Video improves learning in higher education: A systematic review. *Review of Educational Research, 91*(2), 204–236.
van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods, 45*(2), 576–594.
Noroozi, O., Biemans, H., & Mulder, M. (2016). Relations between scripted online peer feedback processes and quality of written argumentative essay. *The Internet and Higher Education, 31*, 20–31.
Ovosi, J. O., Ibrahim, M. S., & Bello-Ovosi, B. O. (2017). Randomized controlled trials: Ethical and scientific issues in the choice of placebo or active control. *Annals of African Medicine, 16*(3), 97–100. https://doi.org/10.4103/aam.aam_211_16
Pai, H. -C. (2015). The effect of a self-reflection and insight program on the nursing competence of nursing students: A longitudinal study. *Journal of Professional Nursing, 31*(5), 424–431.
Panadero, E. (2016). is it safe? Social, interpersonal, and human effects of peer assessment: A review and future directions. *In G. T. L. Brown & L. R. Harris (Eds.), Handbook of human and social conditions in assessment*, 247–266. Routledge.
Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology, 8*(422). https://doi.org/10.3389/fpsyg.2017.00422
Panadero, E., Alonso-Tapia, J., & Huertas, J. -A. (2014). Rubrics vs. self-assessment scripts: effects on first year university students' self-regulation and performance. *Journal for the Study of Education and Development, 31*(1), 149–163.
Panadero, E., Broadbent, J., Boud, D., & Lodge, J. M. (2019). Using formative assessment to influence self- and co-regulated learning: The role of evaluative judgement. *European Journal of Psychology of Education, 34*(3), 535–557. https://doi.org/10.1007/s10212-018-0407-8
Panadero, E., Brown, G. T. L., & 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., Jonsson, A., & Alqassab, M. (2018). Providing formative peer feedback: What do we know? In A. A. Lipnevich, & J. K. Smith (Eds.), *The Cambridge handbook of instructional feedback*. Cambridge University Press. https://doi.org/10.1017/9781316832134.020.
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., Lipnevich, A. A., & Broadbent, J. (2019). Turning self-assessment into selffeedback. In D. Boud, M. D. Henderson, R. Ajjawi, & E. Molloy (Eds.), *The impact of feedback in higher education: Improving assessment outcomes for learners*. Springer.
Panadero, E., Romero, M., & Strijbos, J. W. (2013). The impact of a rubric and friendship on peer assessment: Effects on construct validity, performance, and perceptions of fairness and comfort. *Studies In Educational Evaluation, 39*(4), 195–203.
Papanthymou, A., & Darra, M. (2019). The contribution of learner self-assessment for improvement of learning and teaching process: A review. *Journal of Education and Learning, 8*(1), 48–64. https://doi.org/10.5539/jel.v8n1p48
Pistone, I., Beckman, U., Eriksson, E., Lagerlöf, H., & Sager, M. (2019). The effects of educational interventions on suicide: A systematic review and meta-analysis. *International Journal of Social Psychiatry, 65*(5), 399–412.
Pond, K., Ul-Haq, R., & Wade, W. (1995). Peer review: A precursor to peer assessment. *Innovations in Education and Training International, 32*(4), 314–323.
Ratminingsih, N. M., Marhaeni, A. A. I. N., & Vigayanti, L. P. D. (2018). Self-assessment: The effect on students' independence and writing competence. *International Journal of Instruction, 11*(3), 277–290.
Roehling, M. V., Pichler, S., & Bruce, T. A. (2013). Moderators of the effect of weight on job-related outcomes: A meta-analysis of experimental studies. *Journal of Applied Social Psychology, 43*, 237–252.
Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68.
Sadler, P. M., & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment, 11*(1), 1–31. https://doi.org/10.1207/s15326977ea1101_1
Salanti, G. (2012). Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis. *Many names, many benefits, many concerns for the next generation evidence synthesis tool. Research Synthesis Methods, 3*(2), 80–97.
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), 1049–1066. https://doi.org/10.1037/edu0000190
Sebba, J., Crick, R., Yu, G., Lawson, H., Harlen, W., & Durant, K. (2008). *Systematic review of research evidence of the impact on students in secondary schools of self and peer assessment. I. Research Evidence in Education Library series*. EPPI-Centre, Social Science Research Unit, Institute of Education, University of London. http://eppi.ioe.ac.uk/cms/Default.aspx.
Shen, B., Bai, & Xue, W. (2020). The effects of peer assessment on learner autonomy: An empirical study in a Chinese college English writing class. *Studies In Educational Evaluation, 64*. https://doi.org/10.1016/j.stueduc.2019.100821
Stellmack, M. A., Keenan, N. K., Sandidge, R. R., Sippl, A. L., & Konheim-Kalkstein, Y. L. (2012). Review, revise, and resubmit: The effects of self-critique, peer review, and instructor feedback on student writing. *Teaching of Psychology, 39*(4), 235–244. https://doi.org/10.1177/0098628312456589
Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E. (2018). Developing evaluative judgement: Enabling students to make decisions about the quality of work. *Higher Education, 76*(3), 467–481.
Tannacito, T., & Tuzi, F. (2002). A comparison of e-response: Two experiences, one conclusion. *Kairos, 7*(3), 1–14.
Thomas, B. H., Ciliska, D., Dobbins, M., & Micucci, S. (2004). A process for systematically reviewing the literature: Providing the research evidence for public health nursing interventions. *Worldviews on Evidence-Based Nursing, 1*, 176–184. https://doi.org/10.1111/j.1524-475X.2004.04006.x
To, J., & Panadero, E. (2019). Peer assessment effects on the self-assessment process of first-year undergraduates. *Assessment & Evaluation in Higher Education, 44*(6), 920–932. https://doi.org/10.1080/02602938.2018.1548559
Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249–276.
Topping, K. (2003). Self and peer assessment in school and university: Reliability, validity and utility. In M. S. R. Segers, F. J. R. C. Dochy, & E. C. Cascallar (Eds.), *Optimising new modes of assessment: In search of qualities and standards* (pp. 55–87). Springer.

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

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48.

Vygotsky, L. S. (1978). *Mind in society*. MIT Press.

Wang, H. H., Chen, H. T., Lin, H. S., & Hong, Z. R. (2017). The effects of college students' positive thinking, learning motivation and self-regulation through a self-reflection intervention in Taiwan. *Higher Education Research and Development, 36*(1), 201–216.

Wanner, T., & Palmer, E. (2018). Formative self-and peer assessment for improved student learning: The crucial factors of design, teacher participation and feedback. *Assessment & Evaluation in Higher Education, 43*(7), 1032–1047. https://doi.org/10.1080/02602938.2018.1427698

Weber, K. E., Gold, B., Prilop, C. N., & Kleinknecht, M. (2018). Promoting pre-service teachers' professional vision of classroom management during practical school training: Effects of a structured online-and video-based self-reflection and feedback intervention. *Teaching and Teacher Education, 76*, 39–49.

Wen, L. M., & Tsai, C.-C. (2008). Online peer assessment in an in service science and mathematics teacher education course. *Teaching in Higher Education, 13*(1), 55–67. https://doi.org/10.1080/13562510701794050

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology, 10*, 3087. https://doi.org/10.3389/fpsyg.2019.03087

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

Yan, Z. (2020). Self-assessment in the process of self-regulated learning and its relationship with academic achievement. *Assessment & Evaluation in Higher Education, 45*(2), 224–238. https://doi.org/10.1080/02602938.2019.1629390

Yan, Z., Chiu, M. M., & Ko, P. Y. (2020). Effects of self-assessment diaries on academic achievement, self-regulation, and motivation. *Assessment in Education: Principles, Policy & Practice, 27*(5), 562–583. https://doi.org/10.1080/0969594X.2020.1827221

Yan, Z., Li, Z., Panadero, E., Yang, M., Yang, L., & Lao, H. (2021). A systematic review on factors influencing teachers' intentions and implementations regarding formative assessment. *Assessment in Education: Principles, Policy & Practice, 28*(3), 228–260. https://doi.org/10.1080/0969594X.2021.1884042

Yan, Z. (2022). *Student self-assessment as a process for learning*. Routledge.

Yan, Z., & Boud, D. (2021). Conceptualising assessment-as-learning. In Z. Yan, & L. Yang (Eds.), *Assessment as learning: Maximising opportunities for student learning and achievement* (pp. 11–24). Routledge.

Yan, Z., & Brown, G. T. L. (2017). A cyclical self-assessment process: Towards a model of how students engage in self-assessment. *Assessment & Evaluation in Higher Education, 42*(8), 1247–1262. https://doi.org/10.1080/02602938.2016.1260091

Yan, Z., & Brown, G. T. L. (2021). Assessment for learning in the Hong Kong assessment reform: A case of policy borrowing. *Studies In Educational Evaluation, 68*, Article 100985. https://doi.org/10.1016/j.stueduc.2021.100985

Yan, Z., & Carless, D. (2021). Self-assessment is about more than self: The enabling role of feedback literacy. *Assessment & Evaluation in Higher Education*. https://doi.org/10.1080/02602938.2021.2001431.

Youde, J. J. (2019). A meta-analysis of the effects of reflective self-assessment on academic achievement in primary and secondary populations [Doctoral dissertation, Seattle Pacific University]. *SPU Repository*. https://digitalcommons.spu.edu/soe_etd/48.

Zheng, L., Zhang, X., & Cui, P. (2020). The role of technology-facilitated peer assessment and supporting strategies: A meta-analysis. *Assessment & Evaluation in Higher Education, 45*(3), 372–386. https://doi.org/10.1080/02602938.2019.1644603

Zimmerman, B. J., & Moylan, A. R. (2009). Self-regulation: Where metacognition and motivation intersect. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *The educational psychology series. Handbook of metacognition in education* (pp. 299–315). Routledge.

van Zundert, M. J., Könings, K. D., Sluijsmans, D. M. A., & van Merriënboer, J. J. G. (2012). Teaching domain-specific skills before peer assessment skills is superior to teaching them simultaneously. *Educational Studies, 38*(5), 541–557. https://doi.org/10.1080/03055698.2012.654920

van Zundert, M., Sluijsmans, D., & van Merriënboer, J. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and Instruction, 20*(4), 270–279. https://doi.org/10.1016/j.learninstruc.2009.08.004