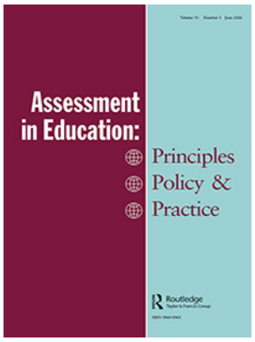

Assessment in Education: Principles, Policy & Practice

ISSN: 0969-594X (Print) 1465-329X (Online) Journal homepage: www.tandfonline.com/journals/caie20

A meta-analytical review of the effects of self-assessment and peer-assessment interventions on self-regulated learning strategies and affective outcomes

Zi Yan, Hongling Lao, Ernesto Panadero & Belén Fernández-Castilla

To cite this article: Zi Yan, Hongling Lao, Ernesto Panadero & Belén Fernández-Castilla (2026) A meta-analytical review of the effects of self-assessment and peer-assessment interventions on self-regulated learning strategies and affective outcomes, Assessment in Education: Principles, Policy & Practice, 33:3, 209-247, DOI: 10.1080/0969594X.2026.2681630

To link to this article: https://doi.org/10.1080/0969594X.2026.2681630

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

View supplementary material

Published online: 07 Jun 2026.

Submit your article to this journal

Article views: 943

View related articles

View Crossmark data

Citing articles: 1 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=caie20

ASSESSMENT IN EDUCATION: PRINCIPLES, POLICY & PRACTICE
2026, VOL. 33, NO. 3, 209–247
https://doi.org/10.1080/0969594X.2026.2681630

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

A meta-analytical review of the effects of self-assessment and peer-assessment interventions on self-regulated learning strategies and affective outcomes

Zi Yan [a], Hongling Lao [b], Ernesto Panadero [c,d] and Belén Fernández-Castilla [e]

[a]Department of Curriculum and Instruction, The Education University of Hong Kong, Hong Kong SAR, China; [b]Education Development Research Center, Shenzhen Institute of Education Sciences, Shenzhen, China; [c]Centre for Assessment Research, Policy and Practice in Education (CARPE, School of Policty and Practice, Institute of Education, Dublin City Univesrity, Dublin, Ireland; [d]Education Regulated Learning and Assessment (ERLA group), Facultad de Educacion y Deporte, Universidad de Deusto, Bilbao, Spain; [e]Department of Methodology of the Social Sciences, Universidad Nacional de Educación a Distancia, Madrid, Spain

ABSTRACT

Self-assessment (SA) and peer-assessment (PA) are increasingly studied for their effects on learning strategies and affective outcomes. This review examined the effect of SA and PA on eight groups of outcome variables: four self-regulated learning strategies (general regulatory strategies, (meta)cognitive strategies, regulation of motivation, regulation of the resources and the environment) and four affective outcomes (attitude, emotion, motivation, and self-efficacy). The analysis included 353 effect sizes from 75 peer-reviewed studies involving K-12 to higher education students, using experimental or quasi-experimental designs. Three-level model analyses showed that SA and PA had positive effects on desirable learning strategies and affective outcomes (Hedges' $g =$ 0.205 to 0.683). Comparatively, SA interventions demonstrate a larger mean effect size on motivation than PA. Additionally, four significant moderators were identified. Overall, the findings provide preliminary evidence that SA and PA support student learning beyond academic performance, with effect strength varying by outcome category and study quality.

ARTICLE HISTORY
Received 5 March 2025
Accepted 25 May 2026

KEYWORDS
Self-assessment; peer-assessment; meta-analysis; learning strategy; affective outcome

Introduction

Self-assessment (SA) and peer-assessment (PA) have drawn increasing attention among researchers and practitioners because they develop students' capacity to evaluate the quality of their work and that of others, monitor learning processes, and address their learning needs (Yan, Lao, et al., 2022). In education literature, SA and PA share common features. For instance, both SA and PA are useful strategies to

CONTACT Zi Yan zyan@eduhk.hk Department of Curriculum and Instruction, The Education University of Hong Kong, 10 Lo Ping Road, Tai Po, N.T, Hong Kong
Supplemental data for this article can be accessed online at https://doi.org/10.1080/0969594X.2026.2681630

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

support student learning and students need to assume an active and reflective role in the process (Panadero, Jonsson, et al., 2016). Empirical studies on various student populations, ranging from K-12 to higher education, consistently support the positive effect of SA and PA on enhancing academic performance, as well documented in meta-analytic reviews by Brown and Harris (2013), Double et al. (2020), Li et al. (2020), and Yan, Lao, et al. (2022). More importantly, the effects of SA and PA on appropriate learning strategies and desirable affective states are likely to have sustainable effects on student learning and, therefore, are crucial to lifelong learning (Yan, 2022). For example, SA and PA have proven to be facilitators of self-regulation, metacognition and effort regulation (Carver & Scheier, 2000; Kim & Ryu, 2013; Panadero & Romero, 2014; Yan, 2020). In addition, the emphasis on student agency in the SA and PA process is likely to influence affective outcomes. As demonstrated in Sitzmann et al.'s (2010) meta-analysis, SA had a stronger association with motivational outcomes (mean effect size of 0.59) than its association with cognitive learning outcomes (mean effect size of 0.34). Additionally, SA and PA can enhance students' learning autonomy (Andrade & Du, 2007; Cassidy, 2007; Hay & Mathers, 2012), intrinsic learning motivation (Kang'ethe, 2017; Mortimer, 1998; Yan et al., 2020), mastery goal orientation (Yan, 2018), psychological safety in the learning environment (van Gennip et al., 2009), and self-efficacy (Alqassab et al., 2018; Kissling & O'Donnell, 2015; Ramdass & Zimmerman, 2008; Yan, Lee, et al., 2022).

Nonetheless, SA and PA exhibit distinct disparities. SA involves evaluating one's own work and demands a high level of self-awareness and self-reflection. Although external input, such as a teacher's feedback, may enhance the quality of the SA process, by nature it is an internal process. In contrast, PA requires students to assess each other's work, assuming the roles of assessor or assessee. PA revolves around interpersonal dynamics, emphasising the importance of mutual trust and psychological safety (Panadero, 2016; van Gennip et al., 2009). PA provides avenues for student interaction, facilitating learning and development by offering a peer's perspective that students can easily grasp. Clearly, SA and PA engage distinct cognitive, motivational, and emotional processes, underscoring the importance of comparing their effects on student learning.

The existing meta-analytic evidence on the effects of SA and PA on learning strategies and affective outcomes is limited in two important and related ways. First, prior meta-analyses have each focused exclusively on either SA or PA and have examined only a narrow subset of outcome variables (Panadero et al., 2017; Zhan et al., 2023). While one meta-analysis has covered both SA and PA together, it focused exclusively on their effects on academic performance (Yan, Lao, et al., 2022), leaving their effects on learning strategies and affective outcomes unaddressed in a joint synthesis. This fragmented approach has produced an incomplete picture of the evidence base. As SRL strategies and affective outcomes are theoretically interrelated (Zimmerman, 2000; Pekrun, 2006), a review that covers only one domain risks obscuring the broader learning ecology in which SA and PA operate and may lead to conclusions that do not generalise across the full range of outcomes that matter for student learning. Second, the absence of a meta-analysis covering both SA and PA on learning strategies and affective outcomes means that a fundamental comparative question remains unanswered: given that SA and PA share common cognitive and metacognitive mechanisms but differ substantially in their interpersonal dynamics and the source of evaluative feedback, do they produce equivalent effects across these outcome domains, or does each have a distinctive profile of strengths? This question is not merely of theoretical interest; it has direct practical implications for educators who must decide which form of student-directed assessment to implement, and under what conditions.

A broader review is therefore not simply a matter of greater comprehensiveness for its own sake. Rather, it is necessary to address these two limitations. By synthesising evidence on both SA and PA across outcome variables spanning SRL strategies and affective outcomes, the current review enables cross-domain and cross-intervention comparisons that a narrower synthesis cannot support. This design allows us to identify not only whether SA and PA are effective, but for *what* they are most effective, and whether their effects are sufficiently similar to justify treating them as a unified category of student-directed assessment or sufficiently different to warrant distinct instructional recommendations. Thus, the current meta-analytic review aims to synthesise the effects of both SA and PA interventions on learning strategies and affective outcomes, covering studies with a wide range of dependent variables, research designs, and student populations from K-12 to higher education.

Self-assessment and peer-assessment

The descriptions of SA and PA differ across studies in the assessment literature (Alqassab et al., 2023; Panadero, Brown, et al., 2016). SA has been conceptualised either as an ability/skill for evaluating one's own performance, an assessment method, or a learning strategy for formative purposes (Yan, 2016). Recent literature tends to highlight the value of SA as a strategy in the learning process. To capture the diversity of self-assessment, Panadero, Brown, et al. (2016) defined it as a wide range of mechanisms and techniques through which students assess and evaluate the qualities of their own learning processes and products. SA may refer to simply assigning a grade/rating to one's own work (i.e. self-grading/self-rating). It may also involve a complex process through which students seek and use feedback, and reflect on their own work against selected criteria to understand their learning needs. Similarly, PA was defined as '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). PA can be carried out in various forms, ranging from simply assigning a grade/rating to the work of peers to processes whereby students evaluate peers' work and offer feedback to one another (Panadero et al., 2018). A recent review by Alqassab et al. (2023) demonstrates the breadth of PA interventions documented in educational settings. The authors identified 10 key design elements of PA, namely: purpose, PA output, group constellation, unit of assessment assessor, unit of assessment assessee, privacy, contact, matching, revision, and scope of involvement. The interplay among these elements gives rise to a diverse array of PA intervention types, each reflecting a distinct configuration of design choices.

Despite the variations of SA and PA forms, an increasing empirical corpus shows that SA and PA are associated with pedagogical values (e.g. Brown & Harris, 2013; Double et al., 2020; Sanchez et al., 2017). In studies that examine the effect of SA and PA in learning environments, SA and PA interventions are usually designed as an action or process to support learning. This is because SA and PA processes can generate insights that could be used to promote students' learning (Panadero et al., 2018; van Zundert et al., 2010; Yan & Brown, 2017). The power of SA and PA serving learning purposes mainly results from students' active and reflective roles in the SA and PA process (Dochy et al., 1999; Topping, 1998; Yan & Carless, 2022). Through actively engaging in SA and PA, students are allowed to generate and use feedback, monitor their own and others' learning progress, and direct future learning. More importantly, in the SA and PA process, students are provided with opportunities to develop desirable long-term capacities, such as self-regulation (Zimmerman, 2000), feedback literacy (Carless & Boud, 2018), and evaluative judgement (Tai et al., 2018).

Categorising self-regulated learning and affective outcomes

Our aim was to investigate all variables related to the effects of SA and PA, with a focus on learning strategies and affective outcomes. The effects on academic performance were not considered because a meta-analysis on that specific topic was recently published (Yan, Lao, et al., 2022). To categorise the different variables from the selected studies we used an inductive approach, concluding that they can be organised around two main groups: self-regulated learning and affective outcomes (see Table 1).

Self-regulated learning is a well-established field in which educational psychologists have invested greatly in the last decades (e.g. Panadero, 2017; Schunk & Greene, 2018). We identified four subgroups of self-regulatory strategies that had been explored in the selected studies. First, *general regulatory strategies*are explored in studies where the main self-regulation strategies, such as self-reflection and strategy use, were dependent variables. Second, *(meta)cognitive strategies* (as used in Panadero et al., 2017) refer to both cognitive strategies (e.g. concentration) and metacognitive strategies (e.g. metacognitive awareness). Third, *regulation of motivation* covers variables such as effort regulation or

Table 1. Categorisation of dependent variables.

Category	Sub-category	Specific dependent variables in primary studies
Self-regulated learning strategies	General regulatory strategies	Self-regulated learning, self-reflection, strategy use, study habit, self-directed learning, independent learning, learner autonomy
	(Meta)cognitive strategies	Assessment ability/skills (e.g. self-testing, number of feedback, self-prediction, self-evaluation/self-correction bias/accuracy, self-evaluation of task performance), cognitive ability (critical thinking, problem-solving, creative tendency, metacognitive awareness/knowledge, information processing, concentration, cognitive presence), test strategies, mental efforts, mental/cognitive load
	Regulation of motivation	Effort regulation
	Regulation of the resources and the environment	Attendance, resource management strategy, teamwork/team behaviour, collaboration, social/peer interaction, communication, quality of peer learning process
Affective outcomes	Attitude	Attitude towards assessment/self-assessment/peer-assessment/peer feedback/teacher feedback/lesson plan/learning/English/project, sense of belongingness, sense of community, positive thinking
	Emotion	Anxiety, stress, (self-)satisfaction
	Motivation	Motivation, motivation to learn (interest, efficacy, mastery-orientation), learning motivation, intrinsic motivation, task value, time spent on task
	Self-efficacy	Self-efficacy in learning, efficacy, academic self-concept, self-efficacy bias, self-confidence in speaking English

time spent on a task. And fourth, *regulation of the resources and the environment* refers to variables, such as attendance, resource management strategy or teamwork.

Affective outcomes are also a crucial area of research in educational psychology usually divided among emotions and motivation (Elliot et al., 2017; Pekrun&Linnenbrink-Garcia, 2014). Here we identified four subgroups of affective outcomes that had been explored for the effects that SA and PA might have on them. First, regarding *attitudinal* variables, researchers have studied if SA and PA influence students' attitudes, such as attitude towards assessment or SA/PA. Second, it has also been explored the effects of SA and PA on *emotions*, such as anxiety, stress, and (self-) satisfaction. Third, researchers have investigated if SA and PA influence *motivation* through constructs such as intrinsic motivation or task value, among others. And fourth, researchers have also examined the effects of SA and PA on *self-efficacy*. Importantly, while some of these categories are also incorporated in self-regulated learning models, we separated them because these four subcategories refer to mostly affect and not to the regulation of learning; therefore, they should be grouped together under affective outcomes. It would have been incorrect to run a meta-analysis with the self-regulated learning and the affective outcomes together because, while related, they are of different natures.

Importantly, our categorisation was subject to expert review. We sent two reputed educational psychology scholars a table including three columns: the main categories, the subcategories, and the variables included in each of them. After several communications, an agreement was reached. Collectively, these eight outcome categories constitute a theoretically grounded operationalisation of student learning that extends beyond immediate academic performance. Within Zimmerman's (2000) cyclical model of self-regulated learning, the regulation of cognition, motivation, and environmental resources are not peripheral to learning but are constitutive of the self-regulatory process itself. Students who develop stronger (meta)cognitive strategies, more effective effort regulation, and better resource management are developing the very capacities that enable sustained and adaptive learning. Similarly, within Pekrun's (2006) control-value theory of achievement emotions, affective states such as self-efficacy, motivation, and emotion are causally implicated in learning processes. They shape the quality of cognitive engagement, the persistence of effort, and the depth of information processing. On these theoretical grounds, the eight outcome categories are not merely associated with learning, but are constitutive of it. It is important to note, however, that the present review provides evidence of the effects of SA and PA on these outcomes as measured at a single point in time. Thus, the claim that improvements in these outcomes translate into sustained long-term learning gains remains a theoretical prediction that requires longitudinal evidence to substantiate.

More specifically, the four SRL strategy subcategories adopted in this review map onto Pintrich (2000, 2004) widely cited framework of self-regulated learning, which identifies four areas of regulation: cognition, motivation/affect, behaviour, and context. Our (meta) cognitive strategies category corresponds to Pintrich's regulation of cognition, capturing both cognitive learning strategies (e.g. concentration, elaboration) and metacognitive activities (e.g. metacognitive awareness, comprehension monitoring). The general regulatory strategies category captures higher-order self-regulation activities (e.g. self-reflection, broad strategy use) that operate across cognitive content and that are conceptually adjacent to Pintrich's regulation of cognition but warrant separate treatment because primary studies frequently report them at this more global level rather than at the level of specific (meta)cognitive strategies. The regulation of motivation category corresponds directly to Pintrich's regulation of motivation/affect, capturing variables such as effort regulation and time-on-task. Finally, the regulation of resources and the environment category combines what Pintrich treats as regulation of behaviour (e.g. help-seeking, persistence) and regulation of context (e.g. environmental management, teamwork), which are tightly intertwined in primary studies in this field and are rarely reported in a way that allows their reliable separation. We acknowledge that combining behaviour and context into a single category is a pragmatic decision; nevertheless, this aggregation is consistent with how these constructs are operationalised in much of the SA and PA empirical literature and with the broader observation that the boundaries between these areas of regulation are functionally porous (Pintrich, 2004; Wolters, 2003). Anchoring the four subcategories in this established taxonomy clarifies that they reflect a theoretically grounded structure rather than an ad hoc construction, while remaining transparent about the points at which our operationalisation departs from a strict one-to-one mapping.

The placement of self-efficacy among the affective outcomes deserves explicit justification, as its conceptual location is genuinely contested in the literature. Self-efficacy can be plausibly placed in at least three different conceptual neighbourhoods: as a motivational belief within self-regulated learning frameworks (Bandura, 1997; Pintrich, 2000, 2004), as part of a broader category of beliefs and perceptions in peer assessment research (Alqassab et al., 2023), or as an affective outcome closely tied to emotional reactions to one's perceived capacity (Pekrun, 2006). Recent meta-analytic work by Panadero et al. (2023) treats self-efficacy as a distinct outcome category alongside, but separate from, self-regulated learning, reflecting precisely this conceptual ambiguity. Our decision to classify self-efficacy among the affective outcomes in the present review reflects two pragmatic considerations rather than a strong theoretical claim that self-efficacy is exclusively or primarily affective in nature. First, in the primary studies included in this review, self-efficacy was typically operationalised through self-report measures of perceived capacity, which align more closely with the perception/belief dimension than with regulatory strategy use. Second, separating self-efficacy from the SRL strategy categories preserves the analytical distinction between students' regulatory behaviours and their motivational – perceptual states, which is an important distinction for interpreting how SA and PA influence different facets of student learning. We acknowledge that this classification is a defensible but not the only defensible choice, and that an alternative structuring (e.g. grouping self-efficacy with motivation under self-regulated learning, following Pintrich, or treating it as a distinct outcome following Panadero et al., 2023) could yield comparable results. This is acknowledged as a limitation of the current categorisation framework.

The effects of self-assessment and peer-assessment on self-regulated learning strategies and affective outcomes

There is theoretical and empirical evidence supporting the effects of SA and PA on self-regulated learning strategies. The primary aim of implementing SA and PA is to generate feedback that could help students optimise their learning strategies which, in turn, enhance their learning outcomes (Meusen-Beekman et al., 2016; Panadero, Jonsson, et al., 2016, 2019; Yan & Carless, 2022). In SA and PA, students evaluate their own and others' work against performance criteria, aiming to identify their learning needs and correct their work correspondingly or give feedback to a peer to change it accordingly. These processes encourage students to reflect on and adjust their current learning strategies or adopt new strategies so as to address the identified learning needs (Brookhart et al., 2004). SA and PA are, therefore, crucial in supporting self-regulated learning (Panadero et al., 2019; Yan, 2022). During this process, students need SA to assess the suitability of their learning goals, track their learning progress in relation to these goals, and reflect on their learning outcomes to determine future improvements (Panadero et al., 2017; Yan, 2020). Additionally, students need PA to acquire external feedback, which aids in fine-tuning their self-regulation process and gaining insights from the self-regulation processes of their peers (Hadwin & Oshige, 2011; Panadero et al., 2019; Yan & Carless, 2022). Moreover, SA or PA itself is a learning process that provides students with plenty of learning opportunities. Scholars (e.g. Boud, 1995; Reinholz, 2016; Topping, 1998; Yan & Brown, 2017) have specified typical actions in the SA and PA process that can serve as learning opportunities, such as determining performance criteria, seeking and providing feedback, reflecting on one's own or others' performance, making and calibrating SA/PA judgements. Engaging in such actions can develop students' general regulatory strategies (Kim & Ryu, 2013; Panadero et al., 2017, 2018; Yan, 2020), (meta)cognitive strategies (Kim, 2009; Özsoy & Ataman, 2009; van den Boom et al., 2007), regulation of motivation (Yan et al., 2020), and regulation of the resources and the environment, such as teamwork and team behaviour (Chang et al., 2012; Dominick et al., 1997).

SA and PA activities directly involve students in the assessment process and, therefore, are likely to influence students' affective outcomes, such as attitude, emotion, motivation, and self-efficacy. Students, when permitted to adopt an active and reflective role in SA and PA, can potentially activate their agency, fostering a heightened sense of autonomy and ownership (Brown & Harris, 2013). Furthermore, their experiences in SA and PA could enhance their self-efficacy regarding the tasks undertaken (Panadero et al., 2016). As proposed by self-determination theory (Ryan & Deci, 2000), this boost in autonomy and self-efficacy may result in an increased intrinsic motivation to learn. Empirical studies have shown that students' engagement in SA and PA can lead to a positive attitude towards learning (Chu, 2013; Wang et al., 2017), less maladaptive emotions, such as stress and anxiety (Pai, 2015; Sivaci, 2020), higher intrinsic motivation (Mortimer, 1998; Yan et al., 2020), and enhanced self-efficacy (Germain et al., 2022; Kissling & O'Donnell, 2015; Panadero, Brown, et al., 2016).

Prior reviews addressing the effect of self-assessment and peer-assessment on learning strategies and affective outcomes

The available meta-analyses on SA and PA explored their validity (i.e. the correlation between SA/PA and other measures of performance) (e.g. SA: Falchikov & Boud, 1989; León et al., 2023; Li & Zhang, 2021; PA: Falchikov & Goldfinch, 2000; Li et al., 2016) and the effects of SA and PA interventions on academic performance (e.g. SA: Brown &

Harris, 2013; Yan, Lao et al., 2022; Youde, 2019; PA: Double et al., 2020; Huisman et al., 2019; Li et al., 2020; Zheng et al., 2020). Concerning the effect of SA and PA on outcomes other than academic performance, narrative reviews tend to support a generally positive effect of SA and PA on self-regulated learning, learning motivation, self-efficacy, engagement, confidence, self-esteem, and quality of student–teacher relationships, but results are mixed (e.g. Brown & Harris, 2013; Panadero, Brown, et al., 2016; Papanthymou & Darra, 2018; Topping, 1998, 2003). Meta-analyses on the effects of SA and PA on learning strategies and affective outcomes are rare. The only exception in the SA field is Panadero et al. (2017) meta-analysis on the effect of SA on self-regulated learning and self-efficacy. They included 19 studies published in 2016 or earlier in four meta-analyses: the effect sizes for three types of SRL (Learning SRL, Negative SRL, and SRL measured qualitatively) were 0.23, 0.65, and 0.43; the effect size on self-efficacy was 0.73. The other exception in the PA field was a recent meta-analysis by Zhan et al. (2023). They reviewed 17 studies published from 2000 to 2022 about the effect of one particular form of PA (i.e. online peer assessment, OPA) on students' higher-order thinking (HOT). They merged various forms of thinking into the concept HOT, which has two sub-categories, i.e. convergent HOT, mainly driven by logical reasoning with established criteria (e.g. critical thinking, reflective thinking, and argumentation) and divergent HOT, which is more spontaneous and non-linear (e.g. creativity and problem-solving). The result showed that the overall effect of OPA on HOT was significant ($g = 0.76$); the effect on convergent HOT ($g = 0.97$) was larger than the effect on divergent HOT ($g = 0.38$). Despite the valuable insights these two meta-analyses provided, their limitations are that they focused exclusively on a narrow sub-group of outcome variables and covered only a small number of studies; therefore, they did not offer a thorough comprehension of the effects of SA and PA. Another meta-analysis by Sitzmann et al. (2010) is relevant in this field. They found that SA's strongest correlations were with two affective evaluation outcomes (i.e. motivation and satisfaction), while the relationship between SA and cognitive learning was moderate. However, this meta-analysis focused on the construct validity of SA, i.e. its correlation with other criterion variables, but not on the effects of SA interventions.

Furthermore, no published meta-analysis attempts to cover both SA and PA in terms of their effects on learning strategies and affective outcomes. There are compelling theoretical and practical reasons to interpret the results of SA and PA studies together and to compare their effects. Theoretically, despite their surface differences, SA and PA engage a functionally similar set of cognitive and metacognitive processes: in both cases, students are required to internalise evaluative criteria, appraise performance against those criteria, and generate or act upon feedback. These shared mechanisms are theoretically grounded in a common framework, providing a principled basis for synthesising and interpreting findings from SA and PA research within a unified analytical framework. Practically, SA and PA differ in important respects. That is, SA is primarily an introspective, self-referential process, whereas PA involves interpersonal dynamics and the integration of external peer perspectives. Comparing their effects can therefore yield actionable insights for educators seeking to determine which form of student-directed assessment is most effective under particular conditions, such as different educational levels, subject domains, or learning objectives. Thus, the current meta-analytic review aims to synthesise the effect of both SA and PA interventions on learning strategies and affective outcomes. To maximise the coverage of the literature, this review intends to include studies with three types of designs, including experimental studies (group-comparison studies with randomisation), quasi-experimental studies (group-comparison studies without randomisation), and studies with a repeated measures design. It also intends to cover studies conducted on mainstream students from K-12 to higher education and from all available years.

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

In addition to synthesising the overall effects, the current meta-analysis examines moderators of intervention effects. Since understanding the conditions under which SA and PA interventions are most effective will inform the design and implementation of future interventions, efforts to identify the moderators are necessary. Due to the limited meta-analyses on this topic, we did not have many references to select potential moderators. Panadero et al. (2017) examined only four possible moderators (i.e. gender, age, SA components, and the agent implementing the intervention) due to the limited number of studies available to investigate more, and identified two significant moderators (i.e. gender and SA components) of the effects of SA interventions on self-efficacy. In the current meta-analysis, we intended to include as many moderators as possible. In particular, our moderators could be categorised into three groups: the characteristics of SA and PA interventions, the characteristics of the sample, and the methodological characteristics of the study.

The first group of moderators are related to the characteristics of SA and PA interventions. It includes SA/PA approach (*quantitative or qualitative assessment*) because the provision of quantitative scores or qualitative comments in SA/PA processes may have different effectson students' subsequent learning (Lipnevich et al., 2016). Past studies have demonstrated that supporting measures used in SA and PA interventions may lead to varied effects. These measures include *the use of online technology* (Ross & Starling, 2008; Wen & Tsai, 2008), *student training* (Brown & Harris, 2013; Li et al., 2016), *teacher training* (Sebba et al., 2008*), teacher feedback* (Han & Xu, 2020; Miller & Geraci, 2011), *and the use of instruments* (Panadero et al., 2013). All these factors were included in the moderator analyses. It should be noted that the selection of these moderators was guided not only by theoretical considerations but also by the practical constraint of reliable codability from primary study reports. Several dimensions that are theoretically relevant to SA and PA design, such as the formative or summative purpose of the assessment, the degree of interactivity or iterativeness of the process, and the directionality or privacy of peer feedback, could not be coded reliably due to insufficient or inconsistent reporting in the primary studies. This constraint is acknowledged as a limitation of the current moderator framework, and is discussed further in the Limitations section.

The second group of moderators are related to the characteristics of the sample. Past studies have revealed gender differences in SA implementation (e.g. Yan, 2018). Review studies also found that SA interventions had a higher effect on girls' self-efficacy than on boys (e.g. Panadero et al., 2017). But these results are not always found. For example, McDonald (2009) reported that male students benefitted more on motivation from SA training. Students' age and educational level are determinants of the accuracy of SA (Boud & Falchikov, 1989; Brown et al., 2015) and PA (Dijks et al., 2018), which, in turn, influence the learning gains from SA and PA. Thus, *gender,mean age*, and *educational level* (primary, secondary, or higher education) were included in the moderator analyses.

Furthermore, the methodological characteristics of studies may affect the results of interventions and, therefore, have to be considered in synthesising multiple studies (Cooper, 2010; McMillan et al., 2013). This is especially crucial in this study as we included the full range of literature with a wide range of different research designs. In particular, we examined whether the effect of SA and PA interventions differs by different methodological characteristics of the studies, including the *type of design, sampling method, source of outcome measures, source and quality of the instrument, report and control for confounders, response rate*, and *attrition rate*.

Aims and research questions of the current review

The aim of this review is to present a comprehensive analysis of the effects of SA and PA on self-regulated learning strategies and affective outcomes. As discussed earlier, we organised these variables into eight distinct categories, comprising four self-regulated learning strategy outcomes (i.e. general regulatory strategies, (meta)cognitive strategies, regulation of motivation, regulation of the resources and the environment) and four affective outcomes (i.e. attitude, emotion, motivation, and self-efficacy). For each category, we conducted a separate meta-analysis, leading to a total of eight meta-analyses. Each meta-analysis was guided by three research questions (RQs):

RQ1: What is the overall effect of SA, PA, and SA+PA (mixed) interventions on students' self-regulated learning strategies and affective outcomes?

RQ2: What is the difference between the effect of SA and PA interventions on students' self-regulated learning strategies and affective outcomes?

RQ3: How do the intervention effects differ by (1) characteristics of the implementation of SA and PA, (2) characteristics of the sample, or (3) methodological characteristics of the study?

Method

This review followed the guidelines for ensuring the quality of a meta-analysis provided by Pigott and Polanin (2020), i.e., (1) comprehensive searching for eligible studies; (2) unbiased screening of identified studies; (3) including important moderators explaining effect size variability; and (4) reliable coding of studies.

Search strategy

ERIC and PsycINFO were chosen as the primary databases for this study because of their wide-ranging coverage of research in the fields of education and psychology, which 'opens the door to more diverse research documents' (Alexander, 2020, p. 12). The initial search started on 6 June 2021 and was completed within the

Table 2. Search terms used in database searches.

Category	Search terms
Self-assessment	'self-assessment' or 'self-evaluation' or 'self-reflection' or 'self-appraisal' or 'self-monitoring' or 'self-grading' or 'self-rating'
Peer-assessment	'peer-assessment' or 'peer-evaluation' or 'peer-feedback' or 'peer-review' or 'peer-comment' or 'peer-monitoring' or 'peer-grading' or 'peer-rating'
Effect	'effect' or 'impact' or 'influence' or 'result' or 'outcome' or 'consequence' or 'contribution' or 'change'

month. A final search was conducted on 5 April 2022 in search of updates in the databases. The search strategy was discussed among the team iteratively and operated mainly by the second author. Search terms employed for this study are listed in Table 2. Each key term of SA and PA was paired with the key terms of effect. In consideration of comprehensiveness, no time limit was set on the publication date. The database search yielded a total of 2,282 hits. Additional 38 hits were identified via alternative sources (e.g. expert reference and forward-searching). Previous related meta-analyses were identified during the literature review and their included study lists were added for further screening. Then, search results from all sources were pooled together. Duplicates were removed by comparing title and author information. After removing duplicates, 1,778 articles remained. Their titles and abstracts were subject to screening based on the inclusion and exclusion criteria.

Inclusion and exclusion criteria

Two sets of inclusion criteria were applied during the screening, including technical and content criteria. Specifically, technical criteria included: (a) selected studies were published in peer-reviewed journals or dissertations; (b) the research design was either experimental, quasi-experimental design with a control group, or repeated measures design; (c) selected studies provided adequate information to compute effect size; (d) selected studies were written in English. Content criteria included: (a) selected studies were empirical studies; (b) selected studies investigated the effects of SA or PA intervention on at least one of the targeted outcome variables (i.e. learning strategies or learning-related affective outcomes); and (c) selected studies focused on mainstream students from kindergarten to higher education. Each study was evaluated one by one based on the aforementioned criteria. It required that all inclusion criteria be met to be kept. In other words, studies were excluded if any one of the inclusion criteria was not met (e.g. studies on students with special needs were excluded because of breaking the content criterion c).

A random sample of 125 studies was selected for a pilot screening process to ensure screening quality. Based on the inclusion and exclusion criteria, the titles, abstracts and full texts of these 125 studies were screened by two researchers independently. A 0.91 inter-rater agreement was reached by comparing their final inclusion decisions. The remaining records were further screened after all disagreements were resolved via discussion and selection criteria were clarified.

The screening resulted in removing 1,252 irrelevant articles. The full texts of the selected 526 articles were further checked. Next, 451 irrelevant articles were removed,

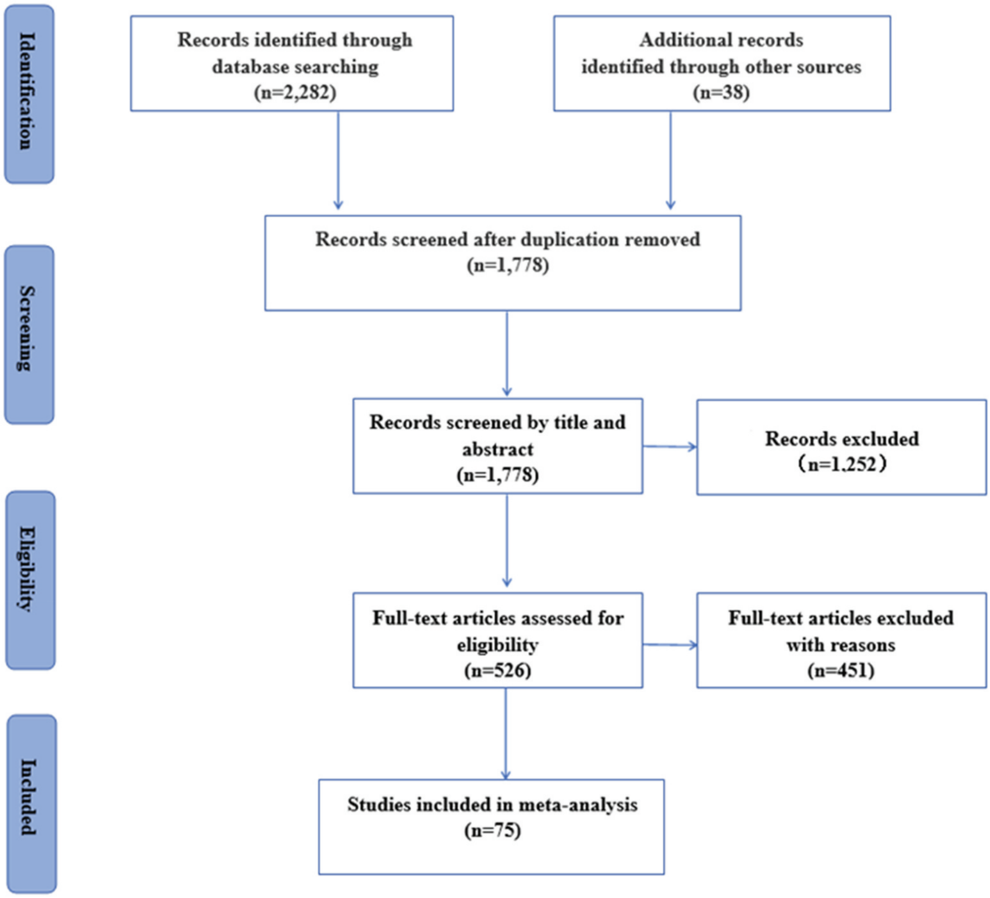

Figure 1. PRISMA flowchart for study identification, screening, and inclusion.

mostly due to a lack of targeted outcome variables. Finally, 75 studies met the inclusion criteria and were included in the analysis (see Table S1 in the Supplementary Materials for a full list of included studies). The process of study identification, screening and inclusion was summarised in the PRISMA flowchart, as shown in Figure 1.

Data extraction

Two sets of information were extracted from each study, including effect size data and moderator data. Effect size data were either reported directly in primary studies or calculated based on the information provided. The effect size was extracted and recorded for each outcome variable separately (e.g. *cognitive and metacognitive strategy*, *resource management strategy*, *affective-attitude*, and *affective-motivation*). Moderator data were further classified into three groups, including the (1) characteristics of SA and PA interventions (i.e. *assessment approach: quantitative, qualitative, or both; the use of online technology: yes or no; teacher training: yes or no; student training: yes or no; teacher feedback: yes or no;* and *the use of assessment instruments: yes or no*), (2) the characteristic of the sample (i.e. *gender; mean age;* and *educational level: pre-school, primary, secondary*, or *higher education;* and *grade*), and (3) the methodological characteristic of studies (i.e. *type of research design: experimental, quasi-experimental, or repeated measures; sampling method: random, systematic*, or *convenient; source of outcome measures: self-reported* [e.g. attitudinal survey] or *assessment/test* [e.g. standardized test scored by teachers]; *instrument source: published, self-developed* or *combined; reporting instrument quality: yes* or *no; report and control for confounders: yes* or *no; response rate;* and *attrition rate*).

To improve data extraction quality, a coding template was developed and pilot-tested with 15 random studies before massive coding. The second and third authors coded the randomly selected studies with the coding template, resulting in an initial .90 inter-rater reliability. In case of disagreement, a consensus was reached via discussion, and the coding template was further clarified and updated accordingly. After the complete coding, additional independent checking was conducted on 20 randomly selected studies, reaching inter-rater reliability of .93.

Study quality

In order to detect inflated overall effects due to the inclusion of low-quality studies that might provide overestimated effects, we quantified the quality of each study using five out of the six items of the Effective Public Health Practice Project (EPHPP) instrument (Thomas et al., 2004). We excluded the item about blinding in the intervention (i.e. the concealment of group allocation) because it is very difficult to blind researchers to the conditions in most educational studies. For the other five items, a label of 'weak', 'moderate', or 'strong' was assigned to each study. The first item refers to whether participants were randomly sampled (strong), systematically sampled (moderate) or conveniently sampled (weak). The second item is related to the design of the study: if there was a control group and students (or full groups) were randomly assigned, then the study was categorised as strong; if the design was longitudinal the study was categorised as moderate; and if there was not a random assignment of the students/group the study was labelled as weak. Regarding the third item, the study was categorised as strong if authors identified potential cofounders and controlled for them, as moderate if authors identified cofounders but did not control for them, and as weak if they did not identify potential confounders. The fourth item is composed of two sub-items. The first sub-item refers to how the data were collected. Studies were categorised as strong if data were collected using a test or an assessment, as moderate if data came from self-report measures, and as weak if this information was not provided. The second sub-item refers to the quality of the instrument used to measure outcomes: the study was strong if the test was standardised, moderate if it was adapted from an already existing instrument, and weak if the instrument was self-developed. Lastly, the fifth item evaluated the dropout and attrition rates. Studies were labelled as strong if the attrition rate was below .20, as moderate if it was between .20 and .40, and as weak if it was above .40. If information about any of the five dimensions was not reported, then a weak label was assigned for that dimension.

We individually examined the impact of each of these quality indicators on the observed effect sizes. Furthermore, for a broader analysis, we calculated a global quality score for each primary study based on the number of items classified as weak. A study with three or more items labelled as weak was eventually classified as a study of weak quality, a study that had two weak items was classified as a study of moderate quality, and a study with just one weak rating was categorised as a strong quality study. Notice that the way the final quality score was calculated differs from the one proposed by the original authors of the EPHPP instrument, who suggest that only studies with zero weak items are labelled as strong quality studies. The reason is that all the studies included in this meta-analysis have at least one weak rating, mainly due to missing information, and applying the original rule without modification would have left no way to differentiate among studies that varied considerably in their methodological quality.

Analyses

First, two global meta-analyses were carried out, pooling together the outcomes related to self-regulated learning (i.e. general regulatory strategies, (meta)cognitive strategies, regulation of motivation, regulation of the resources and the environment) and combining the affective outcomes (i.e. attitude, emotion, motivation, and self-efficacy), respectively. For these two global meta-analyses, only the overall effect sizes, their 95% confidence intervals, variance estimates, and standard errors are reported. Second, independent meta-analyses were performed for the eight types of outcomes: general regulatory strategies, (meta)cognitive strategies, regulation of motivation, regulation of the resources and the environment, attitude, emotion, motivation, and self-efficacy. In each of these eight meta-analyses, more analyses were carried out (e.g. meta-regression and publication bias analyses).

One or several effect sizes were calculated within each study. In studies that used independent group designs, standardised mean differences were obtained by calculating the difference between the mean of the group that received the intervention (i.e. SA, PA or Both) and the mean of the control group, and then dividing this difference by the pooled standard deviation. Positive effect sizes indicate that the intervention group performed better than the control group. When SA and PA or when SA and Mixed were directly compared (no study compared PA with Mixed), a positive effect indicated that SA or Mixed led to a better result than PA. For studies that used repeated measures designs, the standardised mean differences were obtained by subtracting the mean at the post-measure from the mean at the pre-measure divided by the standard deviation of the pre-measure (Morris & DeShon, 2002), so a positive effect size indicates an improvement in the post-test. The sampling variance of this effect size was calculated using the formula of Morris and DeShon (2002), 117, Table 2), which makes standardised mean differences comparable across types of designs (i.e. independent and related groups). However, for repeated measures designs, knowledge about the correlation between pre- and post-measures is needed to compute the sampling variance, but this information was never reported. Therefore, we imputed a reasonable correlation of .50 to calculate the variances. Standardised mean differences (Cohen's *d*) were then converted to Hedges' *g* (Hedges, 1981) to avoid overestimation of the effect when the sample size is small. To interpret the magnitude of the effect sizes, we used the empirical benchmarks provided by Hattie (2012), who reviewed common effect size values observed in studies on educational intervention effectiveness. An effect size smaller than 0.2 was considered small, between 0.2 and 0.4 was considered medium, and above 0.4 was considered large.

It is possible to retrieve several effect sizes within one study, either because several outcomes were measured, or because different instruments were used to measure the same outcome. Effect sizes extracted from the same study are statistically dependent. This dependence has to be taken into account by the meta-analytic statistical model to avoid inflated Type I errors (Becker, 2000). To model dependency among effect sizes, three-level models were applied (Cheung, 2014; van den Noortgate et al., 2013, 2014), which estimate both the variability of effect sizes within studies (Level 2) and between studies (Level 3). To check whether effect sizes were significantly heterogeneous at both levels, likelihood-ratio tests were performed comparing the model fit of the three-level model with the model fit of a model that ignored the between-studies (Level 3) and the within-study variance, respectively (Level 2). If there were fewer than 10 effect sizes available for a given category, a traditional random effect (that is, a two-level model) was carried out (this is the reason why, in some results, a Z statistic is reported instead of a t statistic). To explain the variability in effect sizes across and within studies, moderator variables were introduced one by one in the three-level model, yielding a meta-regression model. To ensure a minimum level of statistical power, statistical results for moderator variables were reported and interpreted only if there were at least 5 effect sizes per category for categorical moderators and at least 10 effect sizes for continuous moderators (Deeks et al., 2019). If this condition regarding the minimum number of effect sizes was not met, the magnitude of the effect was still reported in the tables, but the statistical inference information (t-test statistic and p-value) was removed to avoid wrong statistical interpretations (a significant result based on a small number of effect sizes could be the result of a Type I error). The standard errors from this series of meta-regressions were a *posteriori* corrected using robust variance correction as proposed by Tipton et al. (2019) and Fernández-Castilla et al. (2020). Using robust variance correction on the standard errors prevents inflated Type I errors or false-positive results.

Sensitivity analyses were performed to detect potential outlying effects. To identify them, studentized deleted residuals were obtained for each effect size (Viechtbauer & Cheung, 2010). Effect sizes with studentized deleted residuals below −1.96 or above 1.96 were identified as outliers. These effects and studies were removed, and analyses were repeated without them. In this paper, we report the overall effect obtained with and without outliers for each outcome and each comparison.However, for the moderator analyses, we only report results after outlier removal, as excluding extreme values tends to yield more stable estimates. We acknowledge that this is an analytic decision that may alter the pattern of results, particularly in subsets with few studies. Readers should therefore interpret the moderator results with this caveat in mind.

Finally, several publication bias analyses were performed. First, we generated funnel plots plotting effect sizes against their standard errors and made a visual analysis of the symmetric distribution of these effects in the plot. Second, three-level Egger regression tests were performed to see whether small studies led to larger effects (Fernández-Castilla et al., 2021). If effect sizes were asymmetrically distributed in the funnel plot and the three-level Egger regression test was significant, publication bias could exist. In those cases, we applied the selection model of Vevea and Woods (2005) to get an overall effect size corrected by publication bias. To apply this selection model, effect sizes have to be independent, so we randomly selected one effect size per study. We assumed both a moderate and a severe bias condition. Under moderate bias, the probability that a study reporting a p-value smaller than .05 was published was 1, whereas the probability that a study reporting a p-value larger than .05 was .50. Under a severe bias condition, these probabilities were 1 and .30, respectively.

R studio was used to carry out the analyses. Specifically, main analyses were done with package*metafor* (Viechtbauer, 2010), the robust variance correction was applied with package *clubSandwich* (Pustejovsky, 2022), and the selection model was applied using *weightr*package (Coburn & Vevea, 2019). The dataset with the coding of the studies and the R code that reproduces the results of this paper are available at https://osf.io/kwz7r/?view_only=ae9fc6a0ec904c61be78744498786615.

Results

The current review synthesised 75 independent studies with 353 effect sizes. Forty-two of these studies (including 152 effect sizes) included self-regulatory learning outcomes, and 51 studies (including 201 effect sizes) provided information to calculate effect sizes on affective outcomes (notice that some studies provided results for both types of outcomes). Table 3 includes the overall effect for each comparison and each global outcome,

Table 3. The overall effect, standard errors, variance estimates, and confidence intervals for the overall effect size for each category and each global outcome.

Comparison	Results for SRL outcomes	Results for affective outcomes
SA vs Control	$g = 0.293$, SE = 0.081 ($k = 19$; $m = 72$) $\sigma^2_B = 0.034, I^2_B = 12.89\%$ $\sigma^2_W = 0.167, I^2_W = 64.17\%$ $t = 3.61$, $p = .003$ 95% CI [0.12, 0.47] 4 ES were detected as outliers and removed from these analyses. Including outliers: $g = 0.372$	$g = 0.393$, SE = 0.104 ($k = 25$; $m = 72$) $\sigma^2_B = 0.214, I^2_B = 74.88\%$ $\sigma^2_W = 0.028, I^2_W = 9.80\%$ $t = 3.78$, $p = <.001$ 95% CI [0.18, .61] 13 ES were detected as outliers and removed from these analyses. Including outliers: $g = 0.639$
PA vs control	$g = 0.468$, SE = 0.108 ($k = 18$; $m = 55$) $\sigma^2_B = 0.114, I^2_B = 37.93\%$ $\sigma^2_W = 0.131, I^2_W = 43.29\%$ $t = 4.32$, $p = <.001$ 95% CI [0.24, 0.70] 2 ES were detected as outliers and removed from these analyses. Including outliers: $g = 0.548$	$g = 0.325$, SE = 0.068 ($k = 25$; $m = 104$) $\sigma^2_B = 0.040, I^2_B = 16.45\%$ $\sigma^2_W = 0.159, I^2_W = 65.14\%$ $t = 4.75$, $p = <.001$ 95% CI [0.18, 0.47] 3 ES were detected as outliers and removed from these analyses. Including outliers: $g = 0.381$
SA + PA vs Control	$g = 0.469$, SE = 0.113 ($k = 8$; $m = 18$) $\sigma^2_B = 0.050, I^2_B = 54.84\%$ $\sigma^2_W = 0.000, I^2_W = 0\%$ $t = 4.16$, $p = .007$ 95% CI [0.19, 0.75]	$g = 0.130$, SE = 0.209 ($k = 2$; $m = 4$) $\sigma^2_B = 0.043, I^2_B = 24.23\%$ $Z = 0.63$, $p = .534$ 95% CI [−0.28, 0.54]
SA vs PA	$g = 0.670$, SE = 0.218 ($k = 1$; $m = 1$)	$g = -0.536$, SE = 0.339 ($k = 3$; $m = 3$) $\sigma^2_B = 0.270, I^2_B = 78.67\%$ $Z = 1.58$, $p = .114$ 95% CI [−0.13, 1.20]

Note. SA = self-assessment; PA = Peer assessment, SRL = self-regulatory learning, ES = effect sizes, m = total number of effect sizes, g = overall effect size, SE = standard error, k = number of studies, σ^2_B = between-studies variance, I^2_B = percentage of the total variability that is due to between-studies heterogeneity, σ^2_W = within-study variance, I^2_W = percentage of the total variability that is due to within-study heterogeneity.

its associated standard error, and 95% CIs. An analysis of outliers was carried out before these analyses, and the effect sizes identified as outliers were removed.

For self-regulated learning strategy outcomes, the PA intervention and the SA+PA (mixed) interventions showed higher effectiveness with a large effect size, while the SA intervention showed smaller effectiveness with a medium effect. In terms of affective outcomes, both PA and SA interventions appear to be equally effective with a medium effect.

Moving on to the separate meta-analyses for each of the eight types of outcomes, Table 4 includes the overall effect, standard errors, 95% confidence intervals, and variance estimates for each comparison and sub-outcome (notice that outliers have been previously removed). The next seven sections describe in detail the results for each sub-outcome (except for Regulation of motivation, since there is only one effect size available). The detailed results of the meta-regression analyses are available in Tables S2A-S2G in the Supplementary Materials.

Self-regulated learning strategies: general regulatory strategies

Eleven studies containing 33 effect sizes provided information to calculate a standardised mean difference in the effectiveness of SA versus control on this outcome. Initially, the pooled effect size obtained was 0.101, but two effects were identified as outliers (−2.41 and −2.86) and therefore removed from the analyses. The overall effect (without outliers) was 0.22 (SE = 0.09, $t = 2.45$, $p = .061$, 95% CI [−0.02, 0.45], $\sigma^2_B = 0.000$, $\sigma^2_W = 0.228$), which is a medium effect. Table S2A provides the results of the meta-regressions. Differences between overall effects were detected for the variable 'qualitative assessment'. When there was a qualitative SA (e.g. students write reflective comments on their own performance) the overall effect was significantly larger ($g = .74$) than when there was no qualitative assessment ($g = -0.009$, dif. =0.748, SE = 0.243, $t = 3.08$, $p = .036$). The effect sizes were symmetrically distributed across the funnel plot (see Figure S2A), and the three-level Egger regression test was not statistically significant (B = 1.03, $p = .220$). Therefore, we concluded that publication bias is not likely to affect this set of analyses.

Twenty effect sizes extracted from three studies provided information to calculate a standardised mean difference that represents the effectiveness of the PA intervention against a control group. No outliers were detected. The overall effect was 0.42 (SE = 0.12, $t = 3.52$, $p = .097$, 95% CI [−0.23, 1.07], $\sigma^2_B = 0.019$, $\sigma^2_W = 0.000$), which is a large effect. Table S2A contains the meta-regression results. The mean age was a significant moderator (B = −0.056, SE = 0.004, $t = -14.6$, $p = .042$). The negative sign of the regression estimate indicates that as the mean age of the sample increases, the effectiveness of PA intervention general regulatory strategies decreases. Figure S2A shows the funnel plot for this set of data. The effect sizes were asymmetrically distributed within the figure, which could be indicative of the presence of publication bias. This asymmetry was confirmed by the three-level Egger regression test (B = 25.20, $p < .01$). The selection method of Vevea and Woods (2005) estimated a corrected overall effect of 0.24 (for severe bias) and 0.29 (for moderate bias). In conclusion, these results suggest that the overall effect of PA could be inflated for this outcome.

Table 4. The overall effect size for each comparison and outcome after removing the outlying effect sizes.

	SA vs Control	PA vs Control	SA + PA vs Control	SA vs PA	SA vs SA + PA
General regulatory strategies	$g = 0.220$ ($k = 11$; $m = 33$) 95% CI [−0.015, 0.454] $I_B^2 = 0\%, I_W^2 = 85, 4\%$	$g = 0.421$ ($k = 3$; $m = 20$) 95% CI [−0.23, 1.07]$I_B^2 = 17.17\%, I_W^2 = 0\%$	$g = 0.545$ ($k = 3$; $m = 3$) 95% CI [0.22, 0.87)] $I_B^2 = 0\%$	-($k = 0$)	-($k = 0$)
(meta)cognitive	$g = 0.463$ ($k = 11$; $m = 35$) 95% CI [0.189, 0.763] $I_B^2 = 46.91\%, I_W^2 = 1.13\%$	$g = 0.478$ ($k = 12$; $m = 24$) 95% CI [0.10, 0.85] $I_B^2 = 31.88\%$ $I_W^2 = 60.27\%$	$g = 0.192$ ($k = 4$; $m = 8$) 95% CI [0.09, 0.29)] $I_B^2 = 0\%$	$g = -0.670$ ($k = 1$; $m = 1$)	-($k = 0$)
Regulation of motivation	$g = 0.230$ ($k = 1$; $m = 1$)	-($k = 0$)	-($k = 0$)	-($k = 0$)	-($k = 0$)
Regulation of the resources and the environment	$g = 0.683$ ($k = 2$; $m = 2$) 95% CI [0.33. 1.03)] $I_B^2 = 15.35\%$	$g = 0.444$ ($k = 4$; $m = 11$) 95% CI [−0.138, 1.03] $I_B^2 = 0\%, I_W^2 = 40.64\%$	$g = 0.727$ ($k = 2$; $m = 7$) 95% CI [0.50, 0.95)] $I_B^2 = 0\%$	-($k = 0$)	-($k = 0$)
	SA vs Control	PA vs Control	SA + PA vs Control	SA vs PA	SA vs SA + PA
Attitude	$g = 0.205$ ($k = 4$; $m = 5$) 95% CI [−0.292, 0.702] $I_B^2 = 84.6\%$	$g = 0.308$ ($k = 15$; $m = 78$) 95% CI [0.06, 0.55] $I_B^2 = 18.85\%, I_W^2 = 63.8\%$	$g = 0.154$ ($k = 1$; $m = 1$)	$g = -0.536$ ($k = 3$; $m = 3$) 95% CI [−0.13, 1.20] $I_B^2 = 78.67\%$	$g = -1.263$ ($k = 2$; $m = 2$) 95% CI [0.91, 1.62] $I_B^2 = 28.49\%$
Emotion	$g = 0.615$ ($k = 7$; $m = 13$) 95% CI [−0.041, 1.27] $I_B^2 = 49.87\%, I_W^2 = 41.78$	$g = 0.653$ ($k = 4$; $m = 4$) 95% CI [0.36, 0.94], $I_B^2 = 38.63\%$	$g = -0.273$ ($k = 1$; $m = 1$)	-($k = 0$)	-($k = 0$)
Motivation	$g = 0.601$ ($k = 9$; $m = 12$) 95% CI [0.18, 1.02] $I_B^2 = 85.34\%$, $I_W^2 = 0\%$	$g = 0.229$ ($k = 5$; $m = 11$) 95% CI [−0.05, 0.51] $I_B^2 = 0\%$, $I_W^2 = 72.00\%$	$g = 0.635$ ($k = 1$; $m = 1$)	-($k = 0$)	-($k = 0$)
Self-efficacy	$g = 0.298$ ($k = 17$; $m = 42$) 95% CI [0.13, 0.46] $I_B^2 = 0\%$, $I_W^2 = 66.30\%$	$g = 0.224$ ($k = 5$; $m = 10$) 95% CI [−0.03, 0.48] $I_B^2 = 10.17\%, I_W^2 = 23.2\%$	$g = 0.030$ ($k = 1$; $m = 1$)	-($k = 0$)	-($k = 0$)

Note. m = total number of effect sizes, g = overall effect size, SE = standard error, k = number of studies, σ_B^2 = between-studies variance, I_B^2 = percentage of the total variability that is due to between-studies heterogeneity, σ_W^2 = within-study variance, I_W^2 = percentage of the total variability that is due to within-study heterogeneity.

Only three effect sizes (coming from three different studies) report information to calculate a standardised mean difference of the effect of SA + PA versus a control group for this outcome. Using a standard random-effects model, we obtained an overall effect of 0.55 (SE = 0.17, Z = 3.24, $p = .001$, 95% CI [0.22, .87], $\sigma^2_B = 0.000$). Since the number of effect sizes available was small, we did not conduct meta-regression nor publication bias analyses.

Self-regulated learning strategies: (meta)cognitive strategies

Eleven studies included information to calculate 35 effect sizes that summarise the effectiveness of SA (compared to a control group) on (meta)cognitive outcomes. Initially, the pooled effect size was 0.734, but the analyses of outliers detected three effect sizes that were abnormally large (3.71, 2.90, and −1.42), and therefore were removed from subsequent analyses. As a result, the final, overall effect size was 0.46 (SE = 0.12, t = 3.78, $p = .004$, 95% CI [0.18, 0.76], $\sigma^2_B = 0.115$, $\sigma^2_W = 0.003$), which is a large effect. The only moderator variable that explained part of the variability was whether students received feedback (see Table S2B). The overall effect of those studies where students received feedback was 0.77, and the overall effect of those studies where there was no feedback was 0.23 (dif. =0.54, SE = 0.140, $t = 5.50$, $p = .003$)). Regarding the publication bias analyses, effect sizes seemed to be symmetrically distributed across the funnel plot (Figure S2B), and the three-level Egger regression test was not statistically significant (B = −0.682, $p = .447$).

PA was compared to a control group in 12 studies containing 24 effect sizes. No outliers were detected. The overall effect was 0.48 (SE = 0.17, $t = 2.85$, $p = .018$, 95% CI [0.10, 0.85], $\sigma^2_B = 0.144$, $\sigma^2_W = 0.273$), which is a large effect. None of the moderators explained the variability among effect sizes (Table S2B). The funnel plot (Figure S2B) shows that some small studies provide large effects whereas there are no small studies that report negative effects. Furthermore, the three-level Egger regression test was statistically significant (B = 7.18, $p = .005$), therefore we calculated a corrected overall effect using the selection model of Vevea and Woods (2005). Under the assumption of severe bias, the corrected overall effect was 0.10, and under the assumption of moderate bias the corrected overall effect was 0.23.

Only four studies reported eight effect sizes on the efficacy of SA+PA (mixed) interventions compared to a control group. A traditional random-effects meta-analysis was carried out (due to the small number of effects), yielding an overall effect of 0.19 (SE = 0.05, Z = 3.71, $p = .002$, 95% CI [0.09, 0.29], $\sigma^2_B = 0.000$), which is a small effect. No further analyses were carried out as the number of effect sizes available was small.

Lastly, only one study examined the effectiveness of SA compared to PA. The observed effect in this study was −0.67, meaning that PA was more effective than SA in this specific study.

Self-regulated learning strategies: regulation of motivation

Only one study provided information to compute one effect size about the effect of SA on regulation of motivation compared to a control group. The effect size was 0.23, indicating a medium effect. No further analyses were carried out.

Self-regulated learning strategies: regulation of the resources and the environment

Only two studies reported information to calculate the effectiveness of the SA intervention compared to a control group. A standard random-effects model was applied to combine these two effects. The overall effect was 0.68 (SE = 0.18, 95% CI [0.33, 1.03]). No more analyses were performed for this subset of data due to the small number of effect sizes.

Thirteen effect sizes from four studies were retrieved that summarised the effectiveness of PA versus a control group. Initially, an overall effect size of 0.798 was obtained. However, two effect sizes were labelled as potential outliers (2.74, 2.67) and were removed from the next analyses. The overall effect (after excluding these two effects) was 0.44 (SE = 0.13, $t = 1.97$, $p = .081$, 95% CI [−0.14, 1.03], $\sigma_B^2 = 0.000$, $\sigma_W^2 = 0.105$). None of the study characteristics used as moderator variables significantly explained the variability among effect sizes (see Table S2C). The effect sizes seemed to be symmetrically distributed within the funnel plot (Figure S2C), and the three-level Egger regression test was not statistically significant (B = 2.50, $p = .410$)

Finally, seven effect sizes (from 2 different studies) investigated the effect of SA+PA (mixed) interventions with a control group. A standard random-effects model was applied to combine these effects, obtaining an overall effect of 0.73 (SE = 0.12, $Z = 6.28$, $p < .001$, 95% CI [0.50, 0.95], $\sigma_B^2 = 0.000$). Due to the small number of effect sizes, no moderator analyses nor publication bias were carried out.

Affective outcomes: attitude

Only five studies that included 7 effect sizes reported enough information to calculate an effect size on the efficacy of SA against a control group.Initially, the pooled effect size obtained was 0.773, but two of these effect sizes were identified as outliers (2.99 and 3.83) and were removed from the next analyses. Therefore, the final, overall effect, including 5 effect sizes from 4 studies, was 0.205 (SE = 0.256, $Z = 0.807$, $p = .420$, 95% CI [−0.29, 0.70], $\sigma_B^2 = 0.256$). Since there were fewer than 10 effect sizes, meta-regression and publication bias analyses were not conducted.

A total of 82 effect sizes (from 15 studies) provided information to calculate an effect size that summarises the effectiveness of PA against a control group. The overall effect size obtained was 0.305. However, an analysis of outliers detected 4 outlying effects that were removed from subsequent analyses (3.17, 2.42, 2.19, −1.76), so 78 effect sizes (from 15 studies) were finally analysed. The overall effect was 0.308 (SE = 0.087, $t = 3.55$, $p = .005$, 95% CI [0.06, 0.55], $\sigma_B^2 = 0.117$,$\sigma_W^2 = 0.499$). Table S2D provides the results of the meta-regressions. No study characteristic explained part of the variability observed. The funnel plot (Figure S2D) seemed symmetric, but the three-level Egger regression test was statistically significant (B = 3.89, $p < .05$). Therefore, we obtained an overall effect, corrected by the potential presence of publication bias. Under the assumption of severe bias, the corrected overall effect was 0.100, and under the assumption of moderate bias, the overall effect was 0.165.

Only one study provided information to calculate an effect size on the effectiveness of SA+PA (mixed) against a control group. This overall effect was 0.154. Regarding the direct comparison between SA and PA, only three effect sizes were available. The overall effect was −0.536 (SE = 0.339, Z = 1.58, p = .114, 95% CI [−0.13, 1.20], $\sigma_B^2 = 0.270$), indicating the PA interventions worked better than SA interventions. Finally, there were only two effect sizes on the efficacy of SA against SA + PA. The overall effect was −1.263, meaning that SA intervention led to worse results than SA + PA intervention

Affective outcomes: emotion

A total of seven studies (including 15 effect sizes) compared a SA group with a control group in a dependent variable related to emotions. Initially, the pooled effect size obtained was 0.992, but two effect sizes were identified as outliers (4.12 and 3.83), so only 13 effects (from 7 studies) were included in the analyses. Therefore, the overall effect (after removing the outliers) was 0.615 (SE = 0.267, t = 2.31, p = .07, 95% CI [−0.04, 1.27], $\sigma_B^2 = 0.279, \sigma_W^2 = 0.234$), which is considered a large effect size. None of the study variables explained the observed variability (see Table S2E). The funnel plot looked slightly asymmetric (see Figure S2E), which smaller studies yielding larger effect sizes. The three-level Egger regression test was almost statistically significant (B = 4.64, SE = 2.39, Z = 1.94, p = .053). Given the unclear results of the publication bias analyses, we obtained a corrected overall effect (assuming severe and moderate bias) by applying the selection method of Vevea and Woods (2005). Under the assumption of a severe bias, the corrected overall effect would be 0.44, and under the assumption of moderate bias, the overall effect would be 0.54. The overall corrected effect size would still be of medium magnitude.

Four studies (containing one effect size each) explored the efficacy of PA (against a control group) on a dependent variable related to emotions. No outliers were detected. The overall effect obtained after applying an ordinary random-effects model was 0.65 (SE = 0.15, Z = 4.44, p < .001, 95% CI [0.36, 0.94], $\sigma_B^2 = 0.033$), which is a large effect. No moderator nor publication bias analyses were carried out since the total number of effect sizes was smaller than 10.

Only one study provided information about the effectiveness of SA plus PA (against a control group). This study yielded an effect size of −0.273 to calculate the Hedges´ g. No study reported a direct comparison between SA and PA.

Affective outcomes: motivation

A total of 9 studies (including 15 effect sizes) compared a SA group with a control group in a dependent variable related to motivation. Initially, the pooled effect size obtained was 0.74, but 3 out of the 15 effect sizes were identified as outliers (3.52, 2.59, 2.65). So only 12 effects (from 7 studies) were included in the analyses. The overall effect (after removing the outliers) was 0.60 (SE = 0.18. t = 3.31. p = .012. 95% CI [0.18. 1.02]. $\sigma_B^2 = 0.226.\sigma_W^2 = 0.000$), which is considered a large effect size. None of the study variables explained the observed variability (see Table S2F). The effect sizes were symmetrically distributed in the funnel plot (see Figure S2F), and the Egger regression test was not significant. Therefore, publication bias is not likely to exist.

Five studies (containing 11 effect sizes) explored the efficacy of PA (against a control group) on a dependent variable related to motivation. No outliers were detected. The overall effect obtained after applying a three-level model was 0.23 (SE = 0.08, $t = 2.65$. $p = .079$, 95% CI [−0.05. 0.51], $\sigma^2_{\mathrm{B}} = 0.000, \sigma^2_{\mathrm{W}} = 0.164$), which is a medium effect. None of the study variables explained the observed variability (see Table S2F). Effect sizes were symmetrically distributed in the funnel plot (see Figure S2F), and the three-level Egger regression test was not significant. Therefore, it seems that publication bias is not likely to exist.

Only one study provided information about the effectiveness of SA plus PA (against a control group). This study yielded an effect size of 0.635.

Affective outcomes: self-efficacy

A total of 18 studies (including 48 effect sizes) compared a SA group with a control group in a dependent variable related to self-efficacy. Initially, the pooled effect size obtained was 0.668, but six effect sizes were identified as outliers (4.76, 4.69, 3.04, 3.01, 2.71, 2.58); therefore, only 42 effects (from 17 studies) were included in the analyses. The overall effect (after removing the outliers) was 0.30 (SE = 0.074. $t = 4.00$, $p = .002$, 95% CI [0.13. 0.46], $\sigma^2_B = 0.000, \sigma^2_W = 0.126$), which is considered a medium effect size. Table S2G contains the results of the meta-regressions. Only the variable 'quantitative assessment' explained part of the variability observed: when there was a quantitative assessment, the overall effect was significantly larger ($g = 0.45$, $k = 27$) than when there was not a quantitative assessment ($g = 0.09$, $k = 15$, difference = 0.181, SE = 0.072, $t = 2.516$, $p = .038$). The effect sizes were symmetrically distributed in the funnel plot, although two big positive effects (based on small studies) do not have a negative counterpart (Figure S2G). Furthermore, the three-level Egger regression test was significant (B = 1.57, $p = .027$). The selection method of Vevea and Woods (2005) estimated an overall corrected effect size of 0.12 (assuming severe bias) and 0.21 (for moderate bias).

Five studies (containing 10 effect sizes) explored the efficacy of PA (against a control group) on a dependent variable related to motivation. No outliers were detected. The overall effect obtained after applying a three-level model was 0.22 (SE = 0.09, $t = 2.57$, $p = .070$, 95% CI [−0.03, 0.48], $\sigma^2_{\mathrm{B}} = 0.007, \sigma^2_{\mathrm{W}} = 0.016$). None of the study variables explained the observed variability (see Table S2G). Effect sizes were symmetrically distributed in the funnel plot (see Figure S2G), and the three-level Egger regression test was not significant. Therefore, publication bias is not likely to exist.

Only one study provided information about the effectiveness of SA plus PA (against a control group). This study yielded an effect size of 0.030.

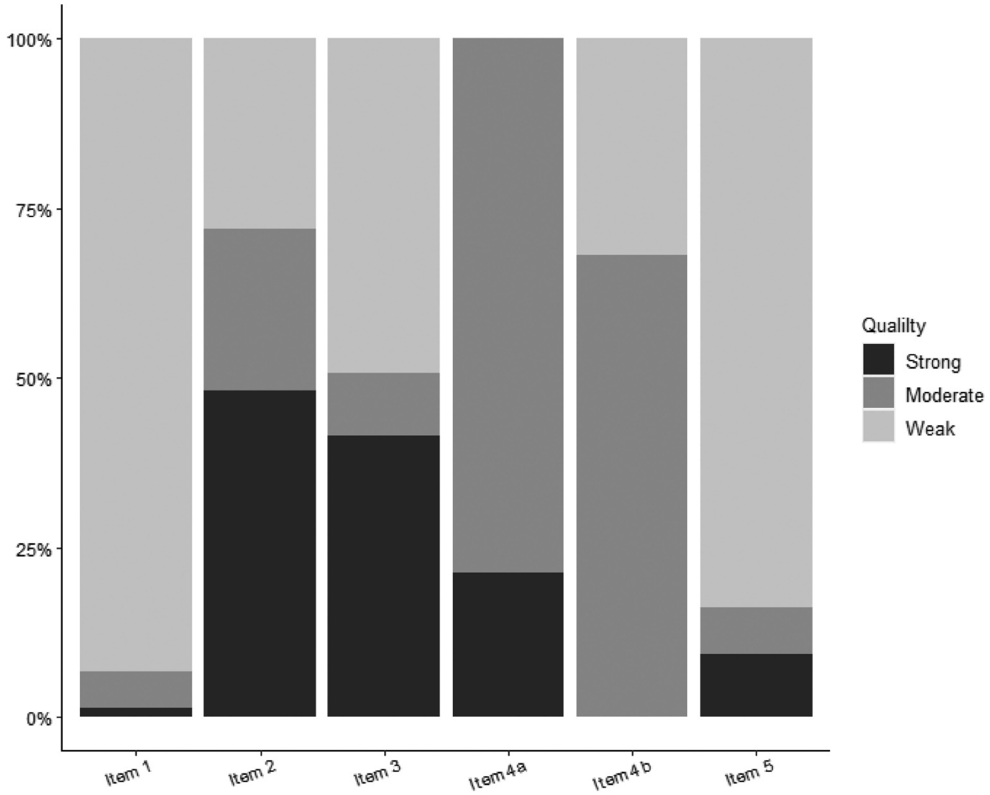

Figure 2. Percentage of studies labelled as strong, moderate or weak quality for each item of the EPHPP instrument. *Notes*. Item 1. Are the individuals selected to participate in the study likely to be representative of the target population? Item 2. Evaluation of the study design. Item 3. Were potential confounders controlled for? Item 4a. Reliability of data collection Item 4b. Evaluation of the instrument used to measure the outcomes Item 5. Evaluation of dropouts and attrition rate. Please refer to the section 'Study Quality' for the coding protocol.

Quality of the studies

The instrument EPHPP categorised three studies as studies with strong quality, 20 studies as moderate, and 52 studies as weak. Figure 2 shows the percentage of studies that were labelled as strong, moderate and weak in each item. As can be seen in that figure, most studies failed to collect a random or systematic sample, and instead used a convenience sample (item 1). Also, none of the studies used standardised tests to measure students' performance, and authors rather applied self-developed or adapted versions of existing surveys (item 4b).

Each quality item was used in the moderator analyses and the significant moderators were reported earlier. In addition, the total quality score of each study was introduced as a moderator for each outcome and each comparison, but it was not significant in any of the analyses.

Discussion

The current meta-analytic review aimed to synthesise the effects of SA and PA interventions on self-regulated learning strategies and affective outcomes. Specifically, we conducted separate meta-analyses for eight outcome variables, including four self-regulated learning strategy outcomes (i.e. general regulatory strategies, (meta)cognitive strategies, regulation of motivation, regulation of the resources and the environment) and four affective outcomes (i.e. attitude, emotion, motivation, and self-efficacy). For each of the eight meta-analyses, we estimated the effects of SA, PA, and SA+PA (mixed) interventions separately, and we compared the effects of SA and PA interventions. We also explored possible moderators of these effects. The current study substantially expands the previous similar meta-analyses (e.g. Panadero et al., 2017; Zhan et al., 2023) by (1) covering a wide range of outcome variables and categorizing them into eight groups; (2) including a much larger number of studies than previous meta-analyses; and (3) including both SA and PA interventions and comparing their effects.

Overall effects of self-assessment, peer-assessment, and the mixed interventions

In general, SA and PA interventions had positive effects across all eight outcome variables (see Tables 3 and 4 for the full numerical details). Several interpretively meaningful patterns emerge from these results. Among the SRL strategy outcomes, (meta)cognitive strategies and regulation of the resources and the environment showed the most consistently strong effects for both SA and PA, suggesting that these two domains are particularly responsive to student-directed assessment interventions. This is theoretically coherent: the core actions involved in SA and PA., e.g., evaluating work against criteria, monitoring progress, and adjusting strategies, map most directly onto cognitive and metacognitive processing and the management of learning resources and effort (Panadero et al., 2017; Zimmerman, 2000). By contrast, the effects on general regulatory strategies and regulation of motivation were more modest, which may reflect the broader and more diffuse nature of these constructs, or the fact that changes in these regulatory habits require more sustained and intensive intervention than a single SA or PA activity can typically provide.

Among the affective outcomes, the most notable finding is the consistently large effects of SA on emotion and motivation. This pattern suggests that SA, as an introspective process directly tied to students' own performance, has a particularly strong capacity to shape how students feel about their learning and how motivated they are to persist. In contrast, PA's effects on affective outcomes were uniformly moderate, which may reflect the more socially mediated and externally oriented nature of the PA process. While PA can foster positive attitudes and self-efficacy through peer interaction, its affective impact may be more variable depending on the quality of the peer relationship and the interpersonal dynamics involved. These differential profiles across outcome domains carry important implications for instructional design, which are discussed further in the Implications section.

The results are broadly consistent with previous theoretical arguments and narrative reviews suggesting that SA and PA have positive effects on learning strategies and affective outcomes (e.g. Brown & Harris, 2013; Panadero, Brown, et al., 2016; Papanthymou & Darra, 2018; Topping, 2003). Moreover, the effects of SA and PA interventions did not differ significantly across gender, educational level, or mean age, with only one exception: the effect of PA on general regulatory strategies decreased when participants' mean age increased. These findings provide preliminary evidence that the value of SA and PA extends beyond academic achievement, and suggest that their effects on learning strategies and affective outcomes merit greater attention in both research and practice. In interpreting these findings, it is useful to distinguish three levels of inferential confidence. First, the positive effects of both SA and PA on (meta)cognitive strategies are among the most consistent findings in this review, supported by a comparatively larger number of effect sizes across both intervention types, and can be stated with reasonable confidence. Second, findings such as the large effects of SA on emotion and motivation are theoretically coherent and consistent with prior work, but rest on fewer studies of predominantly weak quality. Thus, these are best treated as promising patterns warranting further investigation rather than established conclusions. Third, findings for outcome categories with very sparse evidence primarily reveal the underdevelopment of the evidence base and should inform future research priorities rather than serve as a basis for practical recommendations.

Viewing the results across all four elements, i.e. SA, PA, SRL strategies, and affective outcomes, reveals a meaningful pattern that warrants discussion beyond the separate treatment of each outcome domain. SA appears to exert comparatively stronger effects on affective outcomes, particularly emotion ($g = 0.615$) and motivation ($g = 0.601$), whereas PA demonstrates comparatively stronger effects on certain SRL strategy outcomes, particularly general regulatory strategies ($g = 0.421$). This asymmetry is theoretically coherent: SA, as an introspective and self-referential process, is more directly connected to students' internal affective states, while PA, through its interpersonal dynamics and exposure to peers' strategies and perspectives, may be particularly conducive to the development of observable regulatory behaviours. These cross-domain patterns suggest that SA and PA are not interchangeable but rather complementary interventions, each with a distinctive profile of strengths across the SRL strategy and affective outcome domains. Educators seeking to maximise the breadth of student learning gains may therefore benefit from designing instructional environments that incorporate both forms of student-directed assessment in a coordinated manner.

The interplay between self-regulated learning strategies and affective outcomes

While the separate meta-analyses for SRL strategies and affective outcomes are analytically necessary given their conceptual distinctiveness, it is important to recognise that these two domains are not independent in practice. Within Zimmerman's (2000) cyclical SRL model, affective states such as self-efficacy and motivation are not merely outcomes of self-regulatory processes but also serve as inputs that shape subsequent goal-setting, strategic planning, and self-reflection. Conversely, the effective use of regulatory strategies generates experiences of competence that reinforce positive affective states (Bandura, 1997). The findings of the current review are consistent with this interdependence. For instance, the large effects of SA on both (meta)cognitive strategies ($g = 0.463$) and motivation ($g = 0.601$) suggest that engaging in SA may simultaneously strengthen

Table 5. Comparing the overall effect size of SA and PA interventions after removing outlying effect sizes.

	SA vs Control	PA vs Control
General regulatory strategies	$g = 0.220$ ($k = 11$; $m = 33$)	$g = 0.421$ (0.29^*) ($k = 3$; $m = 20$)
(meta)cognitive	$g = 0.463$ ($k = 11$; $m = 35$)	$g = 0.478$ (0.23^*) ($k = 12$; $m = 24$)
Motivation	$g = 0.601$ ($k = 7$; $m = 12$)	$g = 0.229$ ($k = 5$; $m = 11$)
Self-efficacy	$g = 0.298$ (0.21^*) ($k = 17$; $m = 42$)	$g = 0.224$ ($k = 5$; $m = 10$)

Note: *the overall corrected effect size under the assumption of a moderate publication bias.

cognitive engagement and motivational investment – a co-occurrence that is theoretically coherent, as the process of evaluating one's own performance against criteria can foster a heightened sense of agency that activates both more deliberate strategy use and stronger intrinsic motivation (Brown & Harris, 2013). Taken together, these patterns suggest that the full impact of SA and PA on student learning is best understood not by examining SRL strategies and affective outcomes in isolation, but by recognising the potential for a virtuous cycle in which gains in one domain mutually reinforce the other. Future research employing longitudinal designs and mediation analyses would help to illuminate these reciprocal dynamics more explicitly.

Difference between the effects of self-assessment and peer-assessment interventions

An overall comparison between SA and PA (see Table 3) indicates that PA interventions had a larger mean effect size than SA on self-regulated learning strategies, and had a similar mean effect size with SA on affective outcomes. However, the pattern is blurred when we compare SA and PA for each of the eight outcome variables. Table 5 provides information about the relative efficacy of SA and PA interventions on the same outcome variable. Due to the statistical power consideration, only four outcome variables (i.e. general regulatory strategies, (meta)cognitive, motivation, and self-efficacy) with atleast 10 effect sizes for both SA and PA interventions are included in Table 5. The only apparent difference observed from Table 5 is that SA interventions demonstrate a substantially larger mean effect size for the outcome motivation than PA. The differences for the other three outcome variables are not salient, taking into account the publication bias.

The advantage of SA in enhancing motivation could be traced back to Sitzmann et al. (2010) meta-analysis, where they reported that the correlations between SA and affective outcomes were stronger than the connection between SA and cognitive learning. Their findings reasonably led to a conclusion that 'self-assessed knowledge is generally more useful as an indicator of how learners feel about a course than as an indicator of how much they learned from it (p. 180)'. SA is mainly an introspective process that is closely related to one's internal state (Panadero et al., 2013, 2019; Yan & Brown, 2017). Compared to PA, SA is more self-referenced and likely to promote affective reactions as the (internal or external) feedback is directly related to the student's own performance, and there is no direct comparison to peers' work to moderate this affective effect nor the student can talk to peers to downregulate emotions.

However, the advantage of SA in influencing affective outcomes was not supported in the studies that directly compared SA and PA. The difference in effect sizes of SA and PA on attitude in the three effect sizes from three studies (i.e. Ozogul & Sullivan, 2009; Sadeghi & Khonbi, 2015; van Ginkel et al., 2017) was −0.54, favouring PA. The results, however, were not uniform in the three studies. For example, Ozogul and Sullivan (2009) found that the SA group had a more positive attitude towards formative evaluation than the PA group, though the difference was insignificant. In contrast, van Ginkel et al. (2017) reported that PA showed a more salient effect on enhancing students' attitudes towards presenting than SA, yet this difference was also not significant. In Sadeghi and Khonbi's (2015) study, students in the PA condition had a significantly more positive attitude towards assessment than those in the SA condition. There is also one study (i.e. Ozogul & Sullivan, 2009) that directly compared the effects of SA and PA on (meta) cognitive strategies. Based on the descriptive data reported in this study, the standardised mean difference of effect sizes in improving (meta)cognitive strategies (i.e. number of feedback provided) was −0.67, favouring PA. Nevertheless, the small number of studies involving direct comparisons between SA and PA interventions indicates an under-studied area and a worthwhile direction for future research.

Moderators of self-assessment and peer-assessment interventions

Despite the positive mean effect sizes of SA and PA interventions, 25.5% of SA and 18.9% of PA effect sizes were negative, indicating that the effects are not universal and that meaningful heterogeneity exists in the evidence base. The moderator analyses were conducted to identify factors that may account for this variability. However, it is important to interpret the moderator findings with appropriate caution. The analyses identified only four significant moderators overall, and many subgroup comparisons were either null or too sparse for formal inference. This limited explanatory yield reflects, at least in part, the constraints of the available evidence base. That is, the moderator set was restricted to features that could be coded reliably from primary study reports, and several theoretically important dimensions of SA and PA design, such as assessment purpose, and the degree of process interactivity, could not be systematically captured. The present moderator findings should therefore be understood as a partial account of the sources of heterogeneity, identifying some implementation features that appear promising while leaving a substantial portion of the observed variability unresolved. What the current analyses support is the conclusion that certain implementation features are associated with stronger effects and are worth further investigation; they do not yet support a comprehensive account of the conditions under which SA and PA are differentially effective across all eight outcome categories.

We identified three significant moderators of SA interventions. First, SA interventions with a 'qualitative assessment' component (e.g. reflective comments) had a significantly larger effect on general regulatory strategies than those without it. This is probably because qualitative SA provides students with rich opportunities to reflect on their learning and such in-depth reflection may lead to enhanced regulatory strategies. Feedback research often argues that qualitative comments (usually from external parties) are more likely to be supportive of learning than quantitative grades (Lipnevich et al., 2016). Our findings further imply that qualitative feedback generated during SA is also desirable.

Second, teacher feedback plays a crucial role in SA interventions to enhance (meta) cognitive strategies. The SA interventions where students received teacher feedback had a significantly larger effect than those without feedback to students. This finding reiterates that SA is about more than 'self' and external feedback matters in the SA process (Boud, 1999; Butler & Winne, 1995; Panadero et al., 2013, 2019; Yan & Brown, 2017; Yan & Carless, 2022). When SA is implemented as a purely introspective process without external feedback, it is vulnerable to idiosyncratic heuristics and bias (Dunlosky & Rawson, 2012; Joughin et al., 2019), which undermines the association between SA and the gains in desirable cognitive and metacognitive strategies. Thus, students should be encouraged to seek and use teacher feedback to aid their SA.

Third, SA interventions involving quantitative assessment, usually in the form of self-grading or self-rating, resulted in a statistically larger effect on self-efficacy than interventions without quantitative assessment. This result is interesting given that many scholars argue that qualitative SA is more productive than quantitative SA (Boud, 1989; Eva & Regehr, 2008). Our position is that this result should not be interpreted as a counter-argument against the advocacy of qualitative SA. First of all, this factor only moderated the effect of SA on self-efficacy. Quantitative SA evaluation is helpful as it motivates students to play an active role in assessment and enhances their ownership of learning (Nieminen, 2020; Panadero, Brown, et al., 2016). Furthermore, this comparison was between SA interventions with and without quantitative assessment, but not between quantitative and qualitative assessment. Around one-third of SA with quantitative assessment also had a qualitative component. Even for those interventions with quantitative assessment only, students might engage in meaningful qualitative reflection before arriving at a self-grade. In this sense, quantitative SA offers an additional source to enhance students' self-efficacy, but not simply a 'grade guessing' (Boud & Falchikov, 1989; Panadero, Brown, et al., 2016, 2016b).

There is only one significant moderator of PA interventions. Participants with a higher mean age had less improvement in general regulatory strategies. One viable hypothesis is that there is less room for improving regulatory strategies in older students as they have more advanced strategies already. If we look at self-regulated learning research, more experienced students show larger use of regulatory strategies (e.g. Azevedo et al., 2011; Dignath et al., 2008), thus lower chances of improvement.

Comments on the quality of the included studies

Given concerns over the study quality in SA and PA research (e.g. Brown et al., 2015; Panadero & Alqassab, 2019), we provided an overview of the quality of the included studies. The profile of the study quality is not satisfactory – only three studies were rated as strong, 20 studies as moderate, and 52 studies as weak – which may hamper the reliability of their findings. It is crucial to bear in mind that the method used to generate an overall quality score (which relies on the count of quality indicators labelled as weak) may not be ideal. However, considering the generally low quality of the studies, we believe that employing an alternative rule or quality assessment tool would not have led to different conclusions. More importantly, the predominance of weak-quality studies has direct consequences for the interpretive confidence that can be attached to the pooled effect sizes reported throughout this review. The hedged language to describe findings as 'promising' or 'indicative' reflects not only the sparseness of the evidence in certain outcome categories but also the quality limitations of the studies that do exist.

This overview updates the methodological state of the field and informs the directions of future efforts in improving the study quality. There is room for improvement in all five aspects specified in the EPHPP study quality instrument. In particular, future studies should try to improve the study quality through (1) recruiting random or systematic samples instead of convenience samples; and (2) using standardised tests to measure outcome variables.

Importantly, one of the problems with SA and PA is the low quality of the empirical reports. Many articles do not provide sufficient information about the characteristics of the intervention, such as the sample and statistical information, which has been emphasised by some reviews for some time now (e.g. Panadero & Alqassab, 2019). A recent review by Panadero et al. (2023) includes an instrument to report the characteristics of PA designs in an attempt to strengthen incoming PA publications. The instrument we recommend PA scholars to use is accessible here: https://osf.io/5k42z/?view_only=c77740eca9ef44978e1ac47abcaeef7c. Including it in future studies will facilitate the inclusion and analysis of the results in meta-analyses like the present one. We also encourage journal editors and reviewers to pay attention to this issue and request authors to enhance the quality of their reporting in submissions.

Implications for research and practice

The foremost implication for researchers and practitioners is that the value of SA and PA should not be assessed solely on the basis of their effects on academic achievement. The present findings suggest that SA and PA also show positive associations with learning strategies and affective outcomes that are theoretically integral to students' capacity for sustained and self-directed learning (Zimmerman, 2000; Pekrun, 2006). While the causal pathway from these proximal outcomes to long-term learning gains has yet to be established empirically, the theoretical case for treating them as educationally meaningful in their own right is well-grounded, and the present evidence is consistent with that case. The pattern of effect sizes across the eight outcome variables offers more specific guidance for instructional planning. Given that (meta)cognitive strategies and regulation of resources and environment showed the strongest and most consistent effects for both SA and PA, educators aiming to develop students' cognitive and metacognitive capacities are well-positioned to use either form of student-directed assessment. For educators whose primary goal is to strengthen students' motivational investment in learning, SA may be the more targeted choice, given its substantially larger effect on motivation relative to PA. More broadly, the findings suggest that the most comprehensive gains across both SRL strategies and affective outcomes are likely to be achieved through instructional designs that incorporate SA and PA in a coordinated and purposeful sequence, rather than treating them as interchangeable or deploying them in isolation.

Some insights can also be inferred from the moderator analyses. Teacher feedback in SA interventions can enhance their effects on (meta)cognitive strategies, reminding teachers that appropriate scaffolding during SA is vital. Teacher feedback in SA can help students re-calibrate their SA judgements and overcome potential bias. Thus, teachers are advised to provide feedback whenever appropriate, and students should be encouraged to seek and use teacher feedback during SA (for pedagogical recommendations, see Panadero, Jonsson, et al., 2016).

The results show that the type of assessment in SA (qualitative vs. quantitative) matters. On the one hand, including qualitative evaluation in SA enhances the effect of SA interventions on general regulatory strategies, suggesting students should be encouraged to engage in reflective thinking to refine their learning strategies. On the other hand, involving quantitative assessment enhances the effect of SA on self-efficacy, implying that we should rethink the role of self-grading/self-rating. It may be true that 'grade guessing' that does not involve understanding and applying criteria and making meaningful evaluative judgements has limited benefit to student learning (Yan, 2022). However, appropriately designed and implemented self-grading/self-rating may increase students' self-efficacy and offer additional learning opportunities for students to reflect on their learning (Panadero, Brown, et al., 2016). In other words, self-grading/self-rating could be worthwhile if its design can stimulate students' self-efficacy and reflection on their work and if it helps to correct for bias (Panadero, Brown, et al., 2016).

Limitations and future research

This review is not without limitations. First, our intention to include as many outcome variables as possible posed a challenge. That is, how to build a meaningful framework that can cover a variety of variables and, at the same time, recognise the distinctive features of those variables. We proposed the hierarchical categorisation – self-regulated learning strategies (i.e. general regulatory strategies, (meta)cognitive strategies, regulation of motivation, regulation of the resources and the environment) and affective outcomes (i.e. attitude, emotion, motivation, and self-efficacy) – to synthesize the empirical studies. Nevertheless, it does not mean that they are independent variables and could be interpreted in an isolated fashion. As Smith and Ragan (1999) have argued, cognitive objectives often have some affective components, and affective outcomes are usually linked to cognitive goals. We hope this categorisation could be a useful structure to facilitate the interpretation of the results of the current review and call more attention to the effect of SA and PA on learning outcomes other than academic performance. Within this categorisation framework, the placement of self-efficacy among the affective outcomes is a defensible but not the only defensible choice, as discussed in the Categorizing Self-Regulated Learning and Affective Outcomes section. Future meta-analyses with larger evidence bases may be able to disaggregate self-efficacy from other affective outcomes and examine its effects independently, which would allow for a more fine-grained understanding of how SA and PA influence motivational beliefs as distinct from emotional reactions.

Second, the number of studies addressing the effects of SA and PA on many variable outcomes is small. Only one study investigated the effects of SA on regulation of motivation and no study examined PA's effect on it. The available number of effect sizes is less than 10 for SA effect on regulation of the resources and the environment, SA effect on attitude, and PA effect on emotion. For all outcome variables, the available number of effect sizes on the effect of SA+PA (mixed) interventions is less than 10. Also, very few studies directly compared the effect of SA and PA: only one study compared SA and PA on enhancing (meta)cognitive strategies, and three studies compared the effect of SA and PA on attitude. Interpretations of the synthesis results based on these subgroups should be cautious.

Third, indicators of publication bias have been observed for five subgroup comparisons, specifically the general regulatory strategies (PA vs control), (meta)cognitive strategies (PA vs control), attitude (PA vs control), emotion (SA vs control), and self-efficacy (SA vs control). For these subgroup comparisons, studies with small sample sizes reported large and positive effects, whereas no small study reported a large and negative effect. This asymmetry in the funnel plot could lead to biased and inflated pooled effect size estimates. Thus, the interpretation of the results should be cautious and future meta-analyses are encouraged to include more potential missing studies.

Fourth, the moderator framework employed in this review, while covering a range of intervention, sample, and methodological characteristics, does not fully capture the pedagogical heterogeneity of SA and PA as they are enacted in practice. Several dimensions that the literature identifies as consequential for SA and PA design could not be reliably coded from the primary studies, such as the formative or summative purpose of the assessment, the degree of interactivity or iterativeness of the feedback process, and the directionality and privacy of peer feedback. The underreporting of these features in primary studies is itself a methodological concern for the field, as it limits the precision with which meta-analyses can model intervention heterogeneity. As a consequence, the present review operates at a relatively coarse-grained level of SA and PA operationalisation, and the reported average effects should be understood as estimates across a heterogeneous range of pedagogical arrangements rather than as effects of a well-defined, uniform intervention type. Future primary studies should report SA and PA design features in greater detail, and future meta-analyses with access to a richer evidence base should treat the coding of these dimensions as central to the analytical framework rather than peripheral.

Conclusion

The current review provides meta-analytic evidence that SA and PA interventions are associated with positive effects on self-regulated learning strategies and affective outcomes across a range of student populations and educational levels. While the strength and consistency of these effects vary across outcome categories – and while the evidence base for some categories remains limited in size and quality – the overall pattern is broadly consistent with theoretical accounts of how SA and PA support student learning. These findings suggest that the value of SA and PA should not be evaluated solely on the basis of their effects on academic achievement: their potential to support the development of learning strategies and affective outcomes is an equally important consideration for curriculum design and instructional planning. This conclusion is offered with appropriate caution, given the documented limitations of the available evidence, including the sparseness of data in several outcome categories, the generally low methodological quality of the included studies, and the constraints of the moderator framework. The identified significant moderators offer some initial guidance for optimising implementation, though these findings should be treated as preliminary and require replication in higher-quality studies.

Acknowledgments

We would like to thank Miss Chen Jiaying for her contribution to the last round of revisions, including the literature search and review.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The first author was supported by a General Research Fund (GRF) [Project No: EdUHK 18609321] from the Research Grants Council of Hong Kong. The third author was supported via Spanish Ministry of Science, Innovation and Universities under grant [PID2024-156609NB-I00].

Notes on contributors

Zi Yan is a Hong Kong RGC Senior Research Fellow and a Professor in the Department of Curriculum and Instruction at The Education University of Hong Kong. His research focuses on educational assessment, Rasch measurement, and adaptive learning.

Hongling Lao is a researcher at the Education Development Research Center of Shenzhen Institute of Education Sciences, bridging cutting-edge education research, policy, and practice. Her research focuses on large-scale assessment and visualization, innovative pedagogy and assessment literacy, and psychological well-being.

Ernesto Panadero is full professor and Director of Centre for Assessment Research Policy and Practice in Education at the Institute of Education at Dublin City University, and Ikerbasque research professor at the Education Regulated Learning and Assessment team at the Universidad de Deusto (Spain).

Belén Fernández-Castilla is an associate professor in the Department of Methodology of Behavioral Sciences at the Universidad Nacional de Educación a Distancia (UNED). Her research focuses on meta-analytic methods and multilevel modeling.

ORCID

Zi Yan http://orcid.org/0000-0001-9305-884X
Hongling Lao http://orcid.org/0000-0001-9417-8333
Ernesto Panadero http://orcid.org/0000-0003-0859-3616
Belén Fernández-Castilla http://orcid.org/0000-0002-3451-0637

References

Alexander, P. A. (2020). Methodological guidance paper: The art and science of quality systematic reviews. *Review of Educational Research*, *90*(1), 6–23. https://doi.org/10.3102/0034654319854352

Alqassab, M., Strijbos, J. W., Panadero, E., Fernández-Ruiz, J., Warrens, M., & To, J. (2023). A systematic review of peer assessment design elements. *Educational Psychology Review*, *35*(1), 18. https://doi.org/10.1007/s10648-023-09723-7

Alqassab, M., Strijbos, J. W., & Ufer, S. (2018). Preservice mathematics teachers' beliefs about peer feedback, perceptions of their peer feedback message, and emotions as predictors of peer feedback accuracy and comprehension of the learning task. *Assessment and Evaluation in Higher Education*, *44*(1), 139–154. https://doi.org/10.1080/02602938.2018.1485012

Andrade, H., & Du, Y. (2007). Student responses to criteria-referenced self-assessment. *Assessment and Evaluation in Higher Education*, *32*(2), 159–181. https://doi.org/10.1080/02602930600801928

Azevedo, R., Johnson, A., Chauncey, A., & Graesser, A. (2011). Use of hypermedia to assess and convey self-regulated learning. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 102–121). Routledge.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. W.H. Freeman.

Becker, B. J. (2000). Multivariate meta-analysis. In H. E. A. Tinsley & E. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 499–525). Academic Press. https://doi.org/10.1016/B978-012691360-6/50018-5

Boud, D. (1989). The role of self-assessment in student grading. *Assessment and Evaluation in Higher Education*, *14*(1), 20–30. https://doi.org/10.1080/0260293890140103

Boud, D. (1995). *Enhancing learning through self-assessment*. Routledge.

Boud, D. (1999). Avoiding the traps: Seeking good practice in the use of self-assessment and reflection in professional courses. *Social Work Education*, *18*(2), 121–132. https://doi.org/10.1080/02615479911220131

Boud, D., & Falchikov, N. (1989). Quantitative studies of student self-assessment in higher-education: A critical analysis of findings. *Higher Education*, *18*(5), 529–549. https://doi.org/10.1007/BF00138746

Brookhart, S. M., Andolina, M., Zuza, M., & Furman, R. (2004). Minute math: An action research study of student self-assessment. *Educational Studies in Mathematics*, *57*(2), 213–227. https://doi.org/10.1023/B:EDUC.0000049293.55249.d4

Brown, G. T., 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 Publications.

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/10.3102/00346543065003245

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment and Evaluation in Higher Education*, *43*(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

Carver, C. S., & Scheier, M. F. (2000). Scaling back goals and recalibration of the affect system are processes in normal adaptive self-regulation: Understanding "response shift" phenomena. *Social Science and Medicine*, *50*(12), 1715–1722. https://doi.org/10.1016/S0277-9536(99)00412-8

Cassidy, S. (2007). Assessing 'inexperienced' students' ability to self-assess: Exploring links with learning style and academic personal control. *Assessment and Evaluation in Higher Education*, *32*(3), 313–330. https://doi.org/10.1080/02602930600896704

Chang, S. H., Wu, T. C., Kuo, Y. K., & You, L. C. (2012). Project-based learning with an online peer assessment system in a photonics instruction for enhancing LED design skills. *Turkish Online Journal of Educational Technology-TOJET*, *11*(4), 236–246.

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

Chu, R. X. (2013). *Effects of peer feedback on Taiwanese adolescents' English speaking practices and development* [Doctoral dissertationPhD thesis collection]. the University of Edinburgh, Moray House. https://era.ed.ac.uk/handle/1842/8045

Coburn, K. M., & Vevea, J. K. (2019, July 6). *weight: Estimating weight-function models for publication bias*. https://CRAN.R-project.org/package=weightr

Cooper, H. (2010). *Research synthesis and meta-analysis*. Sage.

Deeks, J. J., Higgins, J. P. T., & Altman, D. G. (2019). Analysing data and undertaking meta-analyses. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & V. A. Welch (Eds.), *Cochrane handbook for systematic reviews of interventions* (2nd ed. pp. 241–284). John Wiley & Sons.

Dignath, C., Büttner, G., & Langfeldt, H. (2008). How can primary school students learn self-regulated learning strategies most effectively? A meta-analysis on self-regulation training programmes. *Educational Research Review*, *3*(2), 101–129. https://doi.org/10.1016/j.edurev.2008.02.003

Dijks, M. A., Brummer, L., & Kostons, D. (2018). The anonymous reviewer: The relationship between perceived expertise and the perceptions of peer feedback in higher education. *Assessment and Evaluation in Higher Education*, *43*(8), 1258–1271. https://doi.org/10.1080/02602938.2018.1447645

Dochy, F. J. R. C., 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

Dominick, P. G., Reilly, R. R., & McGourty, J. W. (1997). The effects of peer feedback on team member behavior. *Group & Organization Management*, *22*(4), 508–520. https://doi.org/10.1177/1059601197224006

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

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning & Instruction*, *22*(4), 271–280. https://doi.org/10.1016/j.learninstruc.2011.08.003

Elliot, A. J., Dweck, C. S., & Yeager, D. S. (Eds.). (2017). *Handbook of competence and motivation: Theory and application* (2nd ed.). The Guilford Press.

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

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/10.3102/00346543059004395

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/10.3102/00346543070003287

Fernández-Castilla, B., Aloe, A. M., Declercq, L., Jamshidi, L., Beretvas, S. N., Onghena, P., & van den Noortgate, W. (2020). Estimating outcome-specific effects in meta-analyses of multiple outcomes: A simulation study. *Behavior Research Methods*, *53*(2), 702–717. https://doi.org/10.3758/s13428-020-01459-4

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. https://doi.org/10.1080/00220973.2019.1582470

Germain, L. J., Li, H. H., Wiqas, A., Zahn, L., Stewart, T. M., Hobart, T. R., & Caruso Brown, A. E. (2022). Can peer assessment impact self-efficacy during small-group learning? *PRiMer*, *6*, 6(1). https://doi.org/10.22454/PRiMER.2022.353730

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 and Evaluation in Higher Education*, *45*(5), 680–696. https://doi.org/10.1080/02602938.2019.1689545

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

Hay, M., & Mathers, L. (2012). Designing assessment for autonomous learning. *Practitioner Research in Higher Education*, *6*(2), 95–106.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128. https://doi.org/10.3102/10769986006002107

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 and Evaluation in Higher Education*, *44*(6), 863–880. https://doi.org/10.1080/02602938.2018.1545896

Joughin, G., Boud, D., & Dawson, P. (2019). Threats to student evaluative judgement and their management. *Higher Education Research & Development*, *38*(3), 537–549. https://doi.org/10.1080/07294360.2018.1544227

Kang'ethe, S. M. (2017). Peer assessment as a tool of raising students' morale and motivation: The perceptions of the University of Fort Hare social work students. *International Journal of Educational Sciences*, *6*(3), 407–413. https://doi.org/10.1080/09751122.2014.11890152

Kim, M. (2009). The impact of an elaborated assessee's role in peer assessment. *Assessment and Evaluation in Higher Education*, *34*(1), 105–114. https://doi.org/10.1080/02602930801955960

Kim, M., & Ryu, J. (2013). The development and implementation of a web-based formative peer assessment system for enhancing students' metacognitive awareness and performance in ill-structured tasks. *Educational Technology Research and Development*, *61*, 549–561. https://doi.org/10.1007/s11423-012-9266-1

Kissling, E. M., & O'Donnell, M. E. (2015). Increasing language awareness and self-efficacy of FL students using self-assessment and the ACTFL proficiency guidelines. *Language Awareness*, *24*(4), 283–302. https://doi.org/10.1080/09658416.2015.1099659

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

Li, H., Xiong, Y., Hunter, C. V., Guo, X., & Tywoniw, R. (2020). Does peer assessment promote student learning? A meta-analysis. *Assessment and Evaluation in Higher Education*, *45*(2), 193–211. https://doi.org/10.1080/02602938.2019.1620679

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 and Evaluation in Higher Education*, *41*(2), 245–264. https://doi.org/10.1080/02602938.2014.999746

Li, M., & Zhang, X. (2021). A meta-analysis of self-assessment and language performance in language testing and assessment. *Language Testing*, *38*(2), 189–218. https://doi.org/10.1177/0265532220932481

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.

McDonald, B. (2009). Exploring academic achievement in males trained in self-assessment skills. *International Journal of Primary, Elementary and Early Years Education*, *37*(2), 145–157. https://doi.org/10.1080/03004270802069244

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.

Meusen-Beekman, K. D., Joosten ten Brinke, D., & Boshuizen, H. P. A. (2016). Effects of formative assessments to develop self-regulation among sixth grade students: Results from a randomized controlled intervention. *Studies in Educational Evaluation*, *51*, 126–136. https://doi.org/10.1016/j.stueduc.2016.10.008

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

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), 105–125. https://doi.org/10.1037/1082-989X.7.1.105

Mortimer, J. (1998). Motivating student learning through facilitating independence: Self and peer assessment of reflective practice - an action research project. In S. Brown, S. Armstrong, & G. Thompson (Eds.), *Motivating students* (pp. 173–188). Routledge.

Nieminen, J. H. (2020). Disrupting the power relations of grading in higher education through summative self-assessment. *Teaching in Higher Education*, *27*(7), 892–907. https://doi.org/10.1080/13562517.2020.1753687

Ozogul, G., & Sullivan, H. (2009). Student performance and attitudes under formative evaluation by teacher, self and peer evaluators. *Educational Technology Research and Development*, *57*, 393–410. https://doi.org/10.1007/s11423-007-9052-7

Özsoy, G., & Ataman, A. (2009). The effect of metacognitive strategy training on mathematical problem solving achievement. *International Electronic Journal of Elementary Education*, *1*(2), 67–82.

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. https://doi.org/10.1016/j.profnurs.2015.03.003

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* (pp. 247–266). Routledge.

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

Panadero, E., Alqassab, M., Fernández Ruiz, J., & Ocampo, J. C. (2023). A systematic review on peer assessment: Intrapersonal and interpersonal factors. *Assessment and Evaluation in Higher Education*, *48*(8), 1053–1075. https://doi.org/10.1080/02602938.2023.2164884

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., 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., Jonsson, A., & Strijbos, J. W. (2016). Scaffolding self-regulated learning through self-assessment and peer assessment: Guidelines for classroom implementation. In D. Laveault & L. Allal (Eds.), *Assessment for learning: Meeting the challenge of implementation* (pp. 311–326). Springer.

Panadero, E., Lipnevich, A. A., & Broadbent, J. (2019). Turning self-assessment into self-feedback. In D. Boud, M. D. Henderson, R. Ajjawi, & E. Molloy (Eds.), *The impact of feedback in higher education: Improving assessment outcomes for learners* (pp. 147–163). Springer.

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 peer assessment: Effects on construct validity, performance, and perceptions of fairness and comfort. *Studies in Educational Evaluation*, *39*(4), 195–203. https://doi.org/10.1016/j.stueduc.2013.10.005

Papanthymou, A., & Darra, M. (2018). The contribution of learner self-assessment for improvement of learning and teaching process: A review. *Journal of Education and Learning*, *8*(1), 48. https://doi.org/10.5539/jel.v8n1p48

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review*, *18*, 315–341. https://doi.org/10.1007/s10648-006-9029-9

Pekrun, R., & Linnenbrink-Garcia, L. (Eds.).). (2014). *International handbook of emotions in education*. Routledge.

Pigott, T. D., & Polanin, J. R. (2020). Methodological guidance paper: High-quality meta-analysis in a systematic review. *Review of Educational Research*, *90*(1), 24–46. https://doi.org/10.3102/0034654319877153

Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 451–502). Academic Press. https://doi.org/10.1016/B978-012109890-2/50043-3

Pintrich, P. R. (2004). A conceptual framework for assessing motivation and self-regulated learning in college students. *Educational Psychology Review*, *16*(4), 385–407. https://doi.org/10.1007/s10648-004-0006-x

Pustejovsky, J. (2022, August 15). *Clubsandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections*. https://CRAN.R-project.org/package=clubSandwich

Ramdass, D., & Zimmerman, B. J. (2008). Effects of self-correction strategy training on middle school students' self-efficacy, self-evaluation, and mathematics division learning. *Journal of Advanced Academics*, *20*(1), 18–41. https://doi.org/10.4219/jaa-2008-869

Reinholz, D. L. (2016). The assessment cycle: A model for learning through peer assessment. *Assessment and Evaluation in Higher Education*, *41*(2), 301–315. https://doi.org/10.1080/02602938.2015.1008982

Ross, J. A., & Starling, M. (2008). Self-assessment in a technology-supported environment: The case of grade 9 geography. *Assessment in Education: Principles, Policy & Practice*, *15*(2), 183–199. https://doi.org/10.1080/09695940802164218

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *The American Psychologist*, *55*(1), 68. https://doi.org/10.1037/0003-066X.55.1.68

Sadeghi, K., & Khonbi, Z. A. (2015). Iranian university students' experiences of and attitudes towards alternatives in assessment. *Assessment and Evaluation in Higher Education*, *40*(5), 641–665. https://doi.org/10.1080/02602938.2014.941324

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

Schunk, D. H., & Greene, J. A. (2018). Historical, contemporary, and future perspectives on self-regulated learning and performance. In D. H. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (pp. 1–15). Routledge.

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*. Institute of Education, University of London. http://eppi.ioe.ac.uk/cms/Default.aspx

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 and Education*, *9*(2), 169–191. https://doi.org/10.5465/amle.9.2.zqr169

Sivaci, S. (2020). The effects of peer feedback on writing anxiety levels of pre-service English teachers. *The Reading Matrix: An International Online Journal*, *20*(2), 131–139.

Smith, P., & Ragan, T. J. (1999). *Instructional design*. John Wiley & Sons.

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

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

Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019). A history of meta-regression: Technical, conceptual, and practical developments between 1974 and 2018. *Research Synthesis Methods*, *10*(2), 161–179. https://doi.org/10.1002/jrsm.1338

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research*, *68*(3), 249–276. https://doi.org/10.3102/00346543068003249

Topping, K. J. (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). Kluwer Academic Publishers.

van den Boom, G., Paas, F., & van Merrienboer, J. J. (2007). Effects of elicited reflections combined with tutor or peer feedback on self-regulated learning and learning outcomes. *Learning & Instruction*, *17*(5), 532–548. https://doi.org/10.1016/j.learninstruc.2007.09.003

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. https://doi.org/10.3758/s13428-012-0261-6

van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2014). Meta-analysis of multiple outcomes: A multilevel approach. *Behavior Research Methods*, *47*(4), 1274–1294. https://doi.org/10.3758/s13428-014-0527-2

van Gennip, N. A. E., Segers, M. S. R., & 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. https://doi.org/10.1016/j.edurev.2008.11.002

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. https://doi.org/10.1080/03075079.2015.1117064

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

Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods*, *10*(4), 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. https://doi.org/10.18637/jss.v036.i03

Viechtbauer, W., & Cheung, M. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, *10*(2), 112–125. https://doi.org/10.1002/jrsm.11

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 & Development*, *36*(1), 201–216. https://doi.org/10.1080/07294360.2016.1176999

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

Wolters, C. A. (2003). Regulation of motivation: Evaluating an underemphasized aspect of self-regulated learning. *Educational Psychologist*, *38*(4), 189–205. https://doi.org/10.1207/S15326985EP3804_1

Yan, Z. (2016). The self-assessment practices of Hong Kong secondary students: Findings with a new instrument. *Journal of Applied Measurement*, *17*(3), 335–353.

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 and Evaluation in Higher Education*, *45*(2), 224–238. https://doi.org/10.1080/02602938.2019.1629390

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

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

Yan, Z., & Carless, D. (2022). Self-assessment is about more than self: The enabling role of feedback literacy. *Assessment and Evaluation in Higher Education*, *47*(7), 1116–1128. https://doi.org/10.1080/02602938.2021.2001431

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., Lao, H., Panadero, E., Fernández-Castilla, B., Yang, L., & Yang, M. (2022). Effects of self-assessment and peer-assessment interventions on academic performance: A meta-analysis. *Educational Research Review*, *37*, Article 100484. https://doi.org/10.1016/j.edurev.2022.100484

Yan, Z., Lee, C. K. J., Hui, K. F. S., & Lao, H. L. (2022). Enhancing students' self-efficacy in creativity and learning performance in the context of English learning: The use of self-assessment mind-maps. *Frontiers in Psychology*, *13*, Article 871781. https://doi.org/10.3389/fpsyg.2022.871781

Youde, J. J. (2019). A meta-analysis of the effects of reflective self-assessment on academic achievement in primary and secondary populations. *Education Dissertations, 48*. https://digitalcommons.spu.edu/soe_etd/48

Zhan, Y., Yan, Z., Wan, Z. H., Wang, X., Zeng, Y., Yang, M., & Yang, L. (2023). Effects of online peer assessment on higher-order thinking: A meta-analysis. *British Journal of Educational Technology*, *54*(4), 817–835. https://doi.org/10.1111/bjet.13310

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

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13–40). Academic Press.