Review of Educational Research
October 2021, Vol. 91, No. 5, pp. 761–798
DOI: 10.3102/00346543211019105
Article reuse guidelines: sagepub.com/journals-permissions

When Problem Solving Followed by Instruction Works: Evidence for Productive Failure

Tanmay Sinha and Manu Kapur
ETH Zürich

When learning a new concept, should students engage in problem solving followed by instruction (PS-I) or instruction followed by problem solving (I-PS)? Noting that there is a passionate debate about the design of initial learning, we report evidence from a meta-analysis of 53 studies with 166 comparisons that compared PS-I with I-PS design. Our results showed a significant, moderate effect in favor of PS-I (Hedge's g 0.36 [95% confidence interval 0.20; 0.51]). The effects were even stronger (Hedge's g ranging between 0.37 and 0.58) when PS-I was implemented with high fidelity to the principles of Productive Failure (PF), a subset variant of PS-I design. Students' grade level, intervention time span, and its (quasi-)experimental nature contributed to the efficacy of PS-I over I-PS designs. Contrasting trends were, however, observed for younger age learners (second to fifth graders) and for the learning of domain-general skills, for which effect sizes favored I-PS. Overall, an estimation of true effect sizes after accounting for publication bias suggested a strong effect size favoring PS-I (Hedge's g 0.87).

Keywords: productive failure, direct instruction, preparation for future learning, learning through problem solving

There is a long-standing debate on whether the teaching of a new concept should begin with instruction or problem solving (Tobias & Duffy, 2009). Bringing empirical evidence to bear on this debate is vital for advancing the learning theory as well as practice (Kalyuga & Singh, 2016; Kapur, 2016). This is precisely the aim of our meta-analysis.

Arguments in favor of an *instruction-first approach* (instruction followed by problem solving, or I-PS) are based on decreasing the possibility of student making errors, reducing floundering, and increasing attention to critical and relevant aspects of the domain material (Kirschner et al., 2006). One key underlying assumption is that students are often inadequately equipped to efficiently acquire and consolidate deep learning strategies on their own. Lack of prior knowledge

can lead to time-consuming search through the solution space when students attempt to engage in sensemaking via trial and error, thereby burdening the limited capacity of the working memory. A key recommendation in direct instruction, therefore, is that once targeted domain concepts have been formally introduced and worked examples presented to support solution schema construction, only then are students prepared to be subjected to (un)-guided problem solving in a subsequent phase (Stockard et al., 2018).

Arguments in favor of a *problem-solving first approach* (problem solving followed by instruction, or PS-I) are based on preparing students for future learning (Schwartz & Martin, 2004) by giving them opportunities to notice and encode critical domain features on their own (Loibl et al., 2017). By confronting students with challenging experiences (rather than shrinking the problem-space upfront), their agency (efforts at sensemaking) is emphasized, and learning with germane cognitive load is facilitated. This is achieved by a study design that incorporates an initial exploration phase where students use prior knowledge to develop approximate solutions to novel problems, followed by an instruction phase involving lectures and/or practice (Kapur & Bielaczyc, 2012).

Productive Failure

Productive failure (PF; Kapur, 2008, 2016; Kapur & Bielaczyc, 2012) can be conceived as a subset of PS-I designs that fall under the broader design paradigm of preparation for future learning (PFL; Schwartz & Bransford, 1998). Whereas the problem-solving phase in PF is intentionally designed to result in failure in problem solving, not all problem solving in PS-I is designed to have that feature. See Figure 1 for an illustration depicting this hierarchy. PF comprises an initial generation and exploration phase, affording opportunities for students to activate and differentiate prior and intuitive knowledge, to critique and refine representations and solution methods (RSMs) for solving complex problems. Since these problems are based on concepts students have not formally learned yet, such a problem-solving process very often leads to failure (in relation to a desired goal). In a subsequent consolidation phase, an expert or a teacher builds on student-generated solutions to teach them the targeted concepts. The underlying rationale is to design for failure in the initial learning to minimize failure in the longer term (Kapur, 2016). It must be noted that not all PS-I designs are PF, but only those that follow the design principles of PF as articulated in Kapur and Bielaczyc (2012). We will refer to *PF design fidelity* as the extent to which these criteria are implemented within a PS-I design.

The past decade has seen a growing body of evidence for the efficacy of PF in facilitating conceptual knowledge and transfer (Kapur, 2016; Loibl et al., 2017). Evidence comes not only from quasi-experimental studies conducted in the real ecologies of classrooms (e.g., Hofer et al., 2018; Kapur, 2012; Loibl & Rummel, 2014b) but also from controlled experimental studies (e.g., Kapur, 2014; Newman & DeCaro, 2019; B. Schneider & Blikstein, 2018). Given the explosion in the number of studies that have begun to experimentally investigate the efficacy of different learning designs, we believe our current meta-analytic review (with a focus on comparative interventions between PS-I and I-PS) is very timely. We rigorously examined when, why, and for whom PS-I (and more specifically, PF)

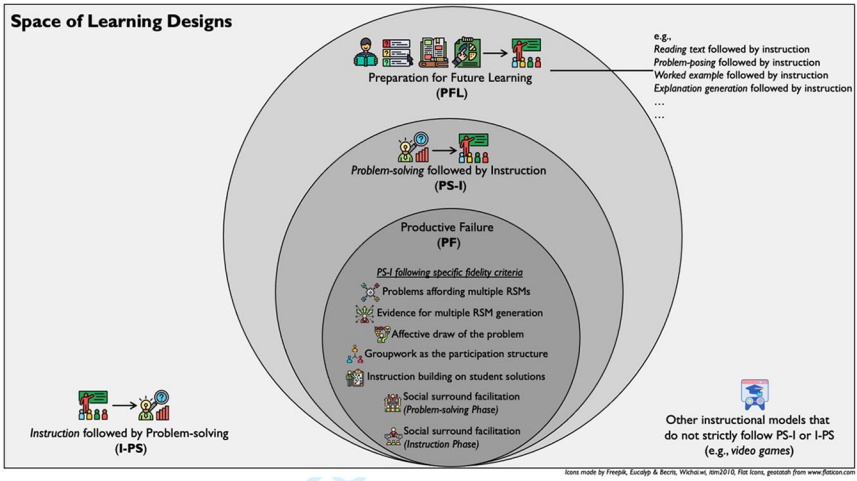

FIGURE 1. *Venn diagram illustrating the hierarchy of PF, PS-I, and PFL learning designs. Here, we depict one category of preparation for future learning (PFL) designs where sensemaking experiences precede instruction. However, more broadly, PFL can be conceived as any experience that prepares students to learn in the future—that learning could occur not just through explicit instruction but also through exploration, practice, and so on. PF = productive failure; PS-I = problem solving followed by instruction.*

works by conducting a systematic meta-analysis of the published literature. Our goal was to understand the extent to which PF design fidelity impacts students' learning, and what salient subgroup differences might be responsible for this effect.

Previous Reviews

Two recent reviews have addressed potential causes of success (Loibl et al., 2017) and failure (Sinha & Kapur, 2019) of designs where problem solving as a preparatory activity precedes instruction. Both reviews (the former based on 34 studies spanning 20 articles, and the latter based on 57 studies spanning 44 articles) have their tradeoffs.

For instance, Loibl et al. (2017) focus on *interrelated cognitive mechanisms* that might explain the positive benefits of approaches implemented based on the PS-I design (e.g., Productive Failure [Kapur & Bielaczyc, 2012], Invent with Contrasting Cases [Schwartz & Martin, 2004]). These mechanisms include intentional activation of relevant prior knowledge, enhancement of learners' awareness of the problem situation and own knowledge gaps, and focused attention on the search for deeper patterns rather than surface characteristics. The *strength* of the Loibl et al. (2017) review lies in that it proposes generalizable cognitive mechanisms that can be systematically tested. However, although Loibl et al. (2017) support their theoretical assumptions about the cognitive mechanisms with some empirical support from the literature, the majority of the studies surveyed were

extremely domain-specific (had mathematics as the learning domain), and were conducted primarily by a handful of authors. Furthermore, Loibl et al. (2017) do not address the question of *when and for whom PS-I works*. In contrast, the surveyed studies in the current meta-analysis come from a diverse demographic range, which affords the opportunity to make more generalizable claims about PS-I efficacy.

The more recent Sinha and Kapur (2019) review doubles the surveyed article base compared to Loibl et al. (2017) and shifts the focus from *why PS-I works* to *when and why it does not work*. Several boundary conditions for the efficacy of PS-I over alternative learning designs (including, but not limited to, only I-PS) are identified. These boundary conditions comprise (a) PF fidelity criteria, (b) incoming characteristics of students (e.g., mastery orientation), (c) nature of the problem-solving task (e.g., domain specificity), (d) student solutions during the problem-solving phase and the extent to which they are scaffolded (e.g., usage of relevant induction criteria), and (e) nuances related to the overall learning design (e.g., additional practice activities). The *strength* of the Sinha and Kapur (2019) review is that it provides a consolidated discussion on when PF fidelity criteria, student, and intervention characteristics negatively affect learning outcomes associated. However, there is no attempt to statistically codify and quantify the strength of the desired effects, which is the aim of the current meta-analysis.

Finally, there is one *quantitative meta-analytic* review of PS-I versus I-PS comparisons (Darabi et al., 2018). This review comprises a small number of articles ($n = 12$) up until 2015 and reports an average effect size of 0.43 [95% confidence interval (*CI*) 0.19, 0.68] in favor of PS-I. Apart from the small sample size and lack of a sufficient number of studies reporting negative effect sizes, there are several limitations of this meta-analysis that may have misrepresented the size of the actually reported effect size estimates. For instance, the methodology (a) comprised a basic *two-level random effects model* that is problematic because it only accounts for participants nested within studies and not the dependency between multiple studies in an article (we address this issue in the current review by using a *three-level meta-analytic model* for pooling effect sizes and conducting all follow-up analyses) and (b) averaged effect sizes across different learning outcomes (e.g., conceptual knowledge, transfer) within a study. We address this issue in the current review by treating these outcomes separately. Furthermore, PS-I was equated with PF, which as we will further argue is not accurate.

The Present Review

Compared to previous reviews (Darabi et al., 2018; Loibl et al., 2017; Sinha & Kapur, 2019), we significantly expand on the number of studies/experimental comparisons (also, potential subgroups we look at). This allows establishing clearly the nature and size of effects produced by PS-I interventions. Identifying conditions under which such preparatory effects are best fostered is of considerable theoretical and practical importance. *Theoretically*, clarifying this issue has important implications for our understanding of the explanatory basis for the efficacy of PS-I. On a *practical* level, such knowledge is also relevant to debates and pedagogic recommendations about methods for improving the efficacy of PS-I learning designs that benefit more diverse student populations.

Research Questions

We address the following research questions:

Research Question 1: What is the differential impact of learning designs that flip the sequence of problem solving and instruction (PS-I and I-PS) on outcomes of (a) procedural knowledge and (b) conceptual knowledge and transfer?

Research Question 2: How does the strength of such differential effects vary with the fidelity of PS-I to PF design criteria, student, and intervention characteristics?

Coding Rationale

PF Fidelity Criteria

Since a *key focus* of the meta-analysis is to ferret out implementation features of PF against I-PS, coding for specific PF fidelity criteria within the included PS-I implementations was a critical first step. Drawing on Kapur and Bielaczyc (2012), designing for PF should take into account the activity engaged in by participants, participant structures used to engage with the problem, and the social surround used to frame the problem-solving context. These elements can be further specified into concrete design criteria spanning the two phases of PF (see Method section for details). We briefly motivate the rationale for these criteria here.

The *generation of multiple RSMs* in the problem-solving phase can be considered a proxy for prior knowledge activation, a key cognitive mechanism (M. Schneider & Stern, 2010) explaining the differential benefit of preparatory learning activities (Loibl et al., 2017). In the absence of evidence for students maximally activating their relevant prior knowledge by generating multiple suboptimal solutions, we would not expect them to gain more from a follow-up lecture, than if they did not participate in any preparatory problem-solving.

Affective draw of the problem is an interrelated fidelity criterion that functions to pique students' situational interest (Hidi & Harackiewicz, 2000). It can be instantiated by creating intuitive hooks that engage students in design via contrasting cases, authentic storylines, and situating the problem within interactive learning environments (Schraw et al., 2001). Empirical results within PF, and more generally PS-I, suggest that such an activity design holds high potential to positively impact curiosity and affect (e.g., Glogger-Frey et al., 2015; Lamnina & Chase, 2019; Loibl & Rummel, 2014a; Sinha et al., 2021; Sinha & Kapur, 2021).

Group work, despite posing additional cognitive costs in coordinating problem-solving strategies from one's partner, is rather well-suited for preparatory problem solving. This is because it affords students the opportunity to cue each other's prior knowledge, build on the complementary expertise of a group member, and use the increased memory and problem-solving resources to detect and correct errors when developing multiple RSMs (Nokes-Malach et al., 2015). However, inhibiting processes such as fear of evaluation from the partner (e.g., when expressing an idea) and social loafing (suboptimal task engagement because of the belief a partner will pick up the slack) may also impede the benefits that can be gained via group work. The tradeoff between these competing mechanisms has

been exemplified in prior PS-I work comparing individual and collaborative PS-I (e.g., Mazziotti et al., 2015; Mazziotti et al., 2019; Sears, 2006), and makes a strong case to assess the differential impact of group work, when implemented within PS-I and compared with a flipped I-PS sequence.

The use of group work as a participation structure should go hand in hand with *social surround facilitation during the problem-solving phase*. By enforcing appropriate socio-mathematical norms and focusing on motivational scaffolds to keep students engaged in RSM generation, the negative influence of threats to students' status and respect is mitigated (Cobb, 1995; Sherin, 2000; Thomas & Brown, 2007). Such threats can arise, for instance, from factors including but not limited to task conflict with peers and coping with uncertainty in RSM generation (because of lack of verifiable outcomes).

The last two PF fidelity criteria focus on the instruction phase of PS-I. *Instruction that builds on student-generated solutions* is more likely to make students aware of specific gaps in their reasoning (Loibl & Rummel, 2014a). By explaining why those gaps exist (via comparison and contrast with the canonical solution), students' perception of the relevance of instructional explanations is likely to lead to deeper processing and higher gains from such consolidation and knowledge assembly. Indeed, theoretical (and empirically tested) accounts of problem solving endorse such an explanatory account (Chi, 2000; VanLehn, 1999). To complement such a form of instruction, *social surround facilitation during the instruction phase* via the use of conversational and social interactive skills (such as critiquing, arguing, engaging students in scientific inquiry, etc.) may be posited to be more effective than an expository style of using one-way instructive presentations (Lazonder & Harmsen, 2016). Within PS-I studies, Loibl and Leuders (2018) and Loibl and Leuders (2019), for instance, have found that explicitly prompting students to elaborate on their errors during follow-up instruction improves students' conceptual knowledge (more so than an experimental condition with monologue-dominant teacher discourse, where it is up to students to connect their activated prior knowledge to the newly presented information).

Students' Incoming Characteristics

Students' *grade level* plays an important role in the context of inductive or deductive learning activities, especially when considering individual differences (e.g., self-regulation skills, mastery orientation) that may be less pronounced at lower grade levels. Young learners (e.g., second to fifth graders) may have insufficient prior knowledge about cognitive and metacognitive learning strategies to generate multiple solutions during initial problem solving (Mazziotti et al., 2015). The resulting lack of relevant prior knowledge activation may attribute to null or negative preparatory benefits. However, other empirical work comparing PS-I and I-PS (Belenky & Nokes-Malach, 2012; DeCaro et al., 2015) has suggested that regardless of students' grade level, those with higher incoming mastery orientation might learn equally well with problem-solving-first or instruction-first approaches. This is because the inventing activity in and of itself provides the motivational impetus to learn the targeted concepts (Belenky & Nokes-Malach, 2012). These contrasting empirical conjectures regarding the dependency of

learning within PS-I on students' *grade level* motivates its inclusion as a potential moderator of the observed effect sizes.

The inclusion of where (or, with what student demographic) a study was carried out as a potential moderator of effect size is driven by two assumptions: (a) first, that certain *geographical distributions* might be dominated by formal lectures as the main pedagogical approach (more so than others), and (b) second, that research groups across these geographies may differ in the diversity of theoretical and methodological commitment to PS-I or I-PS learning design. Both these assumptions may in turn influence the learning outcomes.

Finally, ascertaining what prior knowledge (formal or intuitive) students bring into a learning activity is key to designing appropriate guidance to build upon it (Kapur, 2016). Conducting a pretest on topics similar to (or, different from) those targeted in the intervention is an often-used approach to assess prior knowledge. Furthermore, some empirical work provides support for the idea that a pretest covering concepts targeted during the PS-I intervention might already begin to engage students in preparatory learning mechanisms (Newman & DeCaro, 2019), which, in turn, may dilute the differential advantages of PS-I over I-PS (Kapur, 2016). On the other hand, empirical research on the "forward testing" effect in inductive learning situations suggests that testing of studied information can enhance learning and retrieval of new information (see Yang et al., 2018, for an overview of underlying cognitive mechanisms). With respect to the PS-I design then, facilitatory effects of forward (pre-)testing along with learning mechanisms triggered during the initial problem-solving phase might in fact strengthen (rather than weaken) differential advantages of PS-I over I-PS. In summary, with prior work implicating the presence and *nature of pretesting* to differentially influence problem-solving performance within PS-I (and consequently, learning from instruction), we included this variable as a moderator of the observed effect size.

Intervention Characteristics

Intervention type (experimental vs. quasi-experimental) might serve as a critical confounding factor when interpreting the results of our meta-analysis. In conducting efficacy research to show that PS-I is more effective than variants of standard practice (e.g., I-PS), carrying out experimental work necessitates watering down one or both learning designs to make them similar. Although this strengthens causal attribution about external conditions of learning and/or learning theories, there is a risk that experimental studies end up comparing two suboptimal models of instruction that have the sole merit of differing on only one variable. On the other hand, quasi-experimental research does not limit/constrain natural learning affordances in the service of testing whether one design feature is more effective than another for a specific model of instruction. It may therefore be better poised to compare different instructional paradigms and prove theories about internal learning mechanisms. Furthermore, quasi-experimental studies carried in real classroom ecologies resemble everyday educational practice, as opposed to strictly experimental laboratory studies. Prior empirical work (Glogger-Frey et al., 2015; Hsu et al., 2015) has used some exemplar studies to generalize the claim about lack of (truly) experimental work in the PS-I literature, and hence frequent noncomparability of PS-I and PS intervention conditions. This

motivated coding the *intervention type* to assess the generalizability of this claim and understand what potential factors may be driving the observed trend in effect size estimates between experimental and quasi-experimental comparisons. In the 166 comparisons included in the meta-analysis, we, in fact, found a similar number of experimental comparisons ($n = 81$) and quasi-experimental comparisons ($n = 85$).

A related confound is the *length of exposure* students have to the PS-I or I-PS intervention. On the one hand, repeated exposures to PS-I for student populations (for whom I-PS is the norm) may reduce novelty effects of the learning design and result in similar learning outcomes as I-PS over time. Some previous PS-I work (Kim et al., 2015), however, implicates that repeated exposures to PS-I via multiple smaller cycles of problem solving and instruction happening closely together (resulting in lengthier interventions overall) may, in fact, be beneficial. Given the magnitude/diversity of knowledge assembly students need for understanding different conceptual task elements during the follow-up instruction phase, redundant exposure may result in learning outcomes favoring PS-I. As students spend greater time becoming familiar with the expectations and demands of preparatory problem-solving, it is plausible to expect that they would also gain more from the follow-up lecture and outperform I-PS counterparts on post hoc assessments. These competing conjectures motivate the coding *intervention time span* of each included comparison.

To assess the robustness of the overall effect size estimates across diverse domains, the *targeted learning concept* was included as a potential moderator. Finally, based on prior work that implicates (a) the efficacy of PS-I over I-PS to be more likely for conceptual knowledge and transfer (Kapur, 2016; Loibl et al., 2017) and (b) no clear benefits of PS-I for procedural knowledge, that is, similar or lower performance relative to I-PS (Chen & Kalyuga, 2020; Loibl et al., 2017), *learning outcomes* that we coded focus on these.

Method

Search Criteria

Since our analysis focuses on establishing empirical evidence for when, why, and for whom PF works, our search process and the criteria for including and excluding comparisons for this meta-analysis included articles in the Google Scholar databases[1] that

1. cited either of the two seminal PF articles (Kapur, 2008; Kapur & Bielaczyc, 2012) and/or other key follow-up PF articles (Kapur, 2014, 2015, 2016),[2]
2. reported experimental or quasi-experimental comparison between PF and I-PS, and
3. assessed learning outcomes comprising at least one of conceptual knowledge or transfer and, optionally, also procedural knowledge.

Criterion 1 resulted in 1212 articles as of June 26, 2019, with each of the five PF articles contributing 552, 349, 171, 33, and 107 to the pool, respectively. After

TABLE 1

Descriptive characteristics of articles included in the review (n = 166 comparisons)

Categories	Subcategories	# of comparisons (%)
1. Geographical distribution	Europe (Germany, Switzerland, Belgium, Netherlands)	43 (25.9%)
	North America (USA, Canada)	72 (43.4%)
	Asia (Singapore, Taiwan, India, Hong Kong, Saudi Arabia, Japan)	46 (27.7%)
	Australia	5 (3%)
2. Learner grade/ age range	2nd to 5th graders	25 (15.1%)
	6th to 10th graders	75 (45.2%)
	Undergraduates	61 (36.7%)
	Others (postgraduates, professionals)	5 (3%)
3. Targeted concepts	Math (equivalence, geometry, fractions, variance, linear functions, *z*-scores, statistics process control, fair division/ distribution, crypt-arithmetic)	75 (45.2%)
	Basic sciences[a] (physics, chemistry, biology)	47 (28.3%)
	Physics (average speed, density, collision, electricity, mechanics)	36 (21.7%)
	Chemistry (solutions)	3 (1.8%)
	Biology (genetics, plant adaptations)	4 (2.4%)
	Medicine (dental hygiene/surgery, creatinine clearance, radiographs, suturing, biostatistics)	32 (19.3%)
	Domain general skill (control of variables strategy, water jug problems, Rubik's cube)	8 (4.8%)
	Others (psychology [visual system], environmental science [climate change])	4 (2.4%)

[a]Multiple concepts from all subdomains of basic sciences (physics, chemistry, biology) were covered in four experimental comparisons from Fukaya et al. (2019).

cross-checking this list with 20 articles reported in PS-I's recent qualitative review (Loibl et al., 2017), 324 duplicate records were removed. Forty-five of the remaining 908 articles met Criteria 2 and 3. These 45 articles reported 53 studies and comprised 166 experimental comparisons. References to articles for all comparisons can be found in the supplementary materials, available in the online version of this article. Table 1 presents a breakdown of their demographic characteristics, while Figure 2 depicts a PRISMA flowchart (Moher et al., 2009) summarizing the overall process for selecting studies for inclusion in the meta-analysis. The

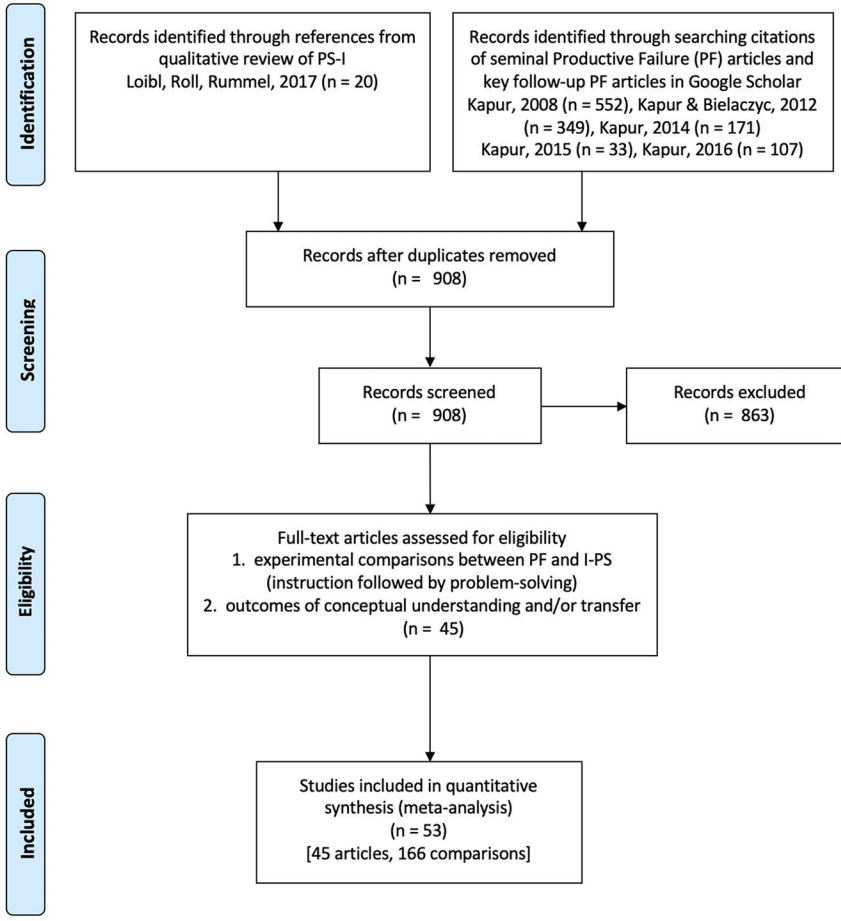

FIGURE 2. *PRISMA flowchart summarizing the overall process for selecting studies for inclusion.*

majority of the studies span a diversity of countries in Europe, North America, and Asia, and cover math concepts for 6th to 10th graders and undergraduates. There is also evidence for PS-I work gradually expanding to populations at the postgraduate and professional levels within other STEM domains like physics and chemistry, non-STEM domains like psychology, and teaching of domain-general skills.

Articles excluded from the meta-analysis broadly fell into the categories of (a) experimental studies (not involving at least one of PS-I or I-PS), (b) observational studies describing student interaction and/or the process of learning, (c) qualitative reviews and syntheses with references to failure (or, negative experiences

more generally), (d) learning analytics and/or methodology-focused work, and (e) qualitative failure-related discourse in application areas beyond education (e.g., music, philosophy, machine design). Please refer to the online supplementary materials for exemplar articles from each of these categories.

Meta-Analysis

With these selected articles, our first goal was to get one overall effect size estimate ($n = 166$).[3] These effect sizes corresponded to the learning outcome of (a) procedural knowledge and (b) conceptual knowledge and/or transfer.[4] Subsequently, these effect sizes were broken down for different relevant subgroupings within the comparisons. Higgins and Green (2011) was used as a reference text to guide all analyses. Implementation was done in *R*, using Harrer et al. (2019) as a technical guide.

Pooling Effect Sizes

A *multilevel meta-analysis model* was used to pool the effect sizes (Assink & Wibbelink, 2016). We explicitly accounted for participants nested within studies, and studies nested within the included articles. The use of such a three-level structure was based on two assumptions. First, studies did not come from the same population; therefore, the deviation in effects between individual studies and the true intervention effect of all studies (due to sampling error) might result in a distribution of true effect sizes (and not just one true effect size). Second, the dependency between effect sizes for multiple studies reported within the same research article might artificially reduce heterogeneity and thus lead to false-positive results; therefore, accounting for this dependency was critical for statistical independence, a core assumption of meta-analytic pooling.

We used Hedge's *g* (bias-corrected standardized mean difference) to assess the overall effect along with 95% confidence intervals (*CI*). The Cochran's *Q* metric (reflecting variance of the distribution of true effect sizes) was used for quantifying statistical heterogeneity in the collected effect size data. However, since Cochran's *Q* can be influenced by the number of studies and their precision (sample size), we also used an additional Higgin's and Thompson's I^2 heterogeneity metric to look at the distribution of variance over the three levels of our meta-analysis model. Such statistical heterogeneity is important to monitor in the collected effect size data since higher values might dilute the confidence we have in our pooled effect.

Subgroup Analysis

To further unpack between-study heterogeneity (that could potentially make effect size estimate less precise) in our three-level meta-analytic model, we looked at a range of subgroup categories that might explain *why* the included studies showed varying results. Essentially, these categories split the data into different subgroups (e.g., fidelity to PF, nature of pretesting, intervention type, etc.). We subsequently analyzed if these subgroups within the studies of our meta-analysis differed in terms of their effects. Subgroups, along with the definition/criteria we used for their coding, are listed below.

1. PF fidelity criteria within the PS-I design (κ = 0.92, 1.0, 0.87, 1.0, 1.0, 0.73, and 0.91). Based on the design principles of PF (Kapur, 2012, 2016) and the nature of information found in the reviewed articles, we were able to code for the following seven criteria:
    a. Problems affording multiple RSMs: *yes* (92.77% comparisons), *no*
    b. Qualitative or quantitative evidence for multiple RSM generation in the article (e.g., if authors tabulated quantity/quality of solutions students generated)[5]: *yes* (59.64% comparisons), *no*
    c. Affective draw of the problem considered (e.g., if the problem design comprised story problems within/outside virtual environments, contrasting cases, simulations, agent-based modeling, etc.): *yes* (61.45% comparisons), *no*
    d. Group work as the participation structure during the problem-solving phase: *yes* (53.01% comparisons), *no*
    e. Instruction building on student solutions: *yes* (44.58% comparisons), *no*
    f. Social surround facilitation during problem-solving phase[6]: *high* (if the intervention explicitly mentioned (considered) more than one of the following subcriteria: (a) the problem-solving phase as a safe space to explore and generate ideas, (b) the social and mathematical norms around the activity (e.g., it is okay not to be able to solve problems as long you try various ways of solving them; highlighting to students that there are multiple solution approaches for the problem solving; setting the expectation that understanding why and under what conditions some solutions are better than others is important); (c) provision of affective/motivational support to persist/keep trying (excluding prompts for cognitive or metacognitive support), and/or if teacher training was held on the PF instructional design prior to the intervention), *low* (only one of these criteria were considered).
    g. Social surround facilitation during instruction phase: *dialogue-dominant* (33.73% comparisons, the teacher asked clarifying questions and/or paraphrased student solutions and/or asked students to elaborate upon each other's ideas and/or engaged students in arguments/conflict and/or used other kinds of facilitation strategies that enhance student engagement), *monologue-dominant* (one-way transmission of information about the concept, expert solutions and/or common student solutions).
2. Students' demographic and/or incoming characteristics (κ = 1.0, 1.0, and 0.84)
    a. Age range (as assessed by their grade level): *2nd to 5th graders*, *6th to 10th graders*, *undergraduates*, *others (postgraduates, professionals)*
    b. Geographical distribution: *Europe*, *North America*, *Asia*, *Australia*
    c. Nature of pretesting[7]: *none* (no pretest administered), *prerequisite* (pretest was administered on prerequisite concepts but not those targeted in the intervention), *targeted* (pretest was administered on concepts similar to those targeted in the intervention)

3. Intervention characteristics ($\kappa = 1.0, 1.0, 1.0, 1.0$)
    a. Type: *experimental* (random assignment to treatment or control groups; groups differ only in terms of the experimental treatment they receive, that is, a flipped sequence of problem solving and instruction), *quasi-experimental* (classroom studies where random assignment is not possible and whole classes are assigned to the treatment or control groups; treatment and control groups differ in more than just the experimental treatment they receive, e.g., different teachers, presence of prompts or accuracy feedback during problem solving, different presentation modality during instruction)
    b. Time span: *a few hours* (within 1 day), *a few days* (>1 day). The intervention length comprises the actual experiment (problem solving and instruction phases), along with any administered pretests and posttests.
    c. Learning concepts targeted: *Math*, *Basic sciences* (*physics, chemistry, biology*), *Medicine*, *Domain-general skill* (e.g., control of variable strategy), *Others* (*psychology, environmental science*)
    d. Learning outcomes assessed: conceptual knowledge, transfer, clubbed outcomes (single posttest assessing conceptual knowledge and transfer), procedural knowledge

For PF design fidelity criteria, the primary authors discussed the application of the criteria to 20% example articles and agreed on the coding criteria, after which the first author (with the help of a research assistant) coded the remaining comparisons. Inter-rater reliability was $\kappa > 0.7$. Disagreements were resolved with discussion. The coding scheme for other relatively more subjective categories such as social surround facilitation was iteratively developed. Interrater reliability $\kappa > 0.7$ was established, after which one rater coded the rest of the data. As before, disagreements were resolved with discussion. Table 2 showcases some examples of comparisons with varying PF design fidelity (Loehr et al., 2014; Loibl & Rummel, 2014a; Song, 2018), along with the coded subgroup category for those comparisons. Overall, in terms of the extent to which PF is representative of approaches to PS-I instruction, the least represented PF fidelity criteria are social surround facilitation (within the problem solving and instruction phases), instruction building on student solutions, and the use of group work as the participation structure. All coded comparisons with subgroup categories and implementation-level details for PS-I and I-PS can be found in the online supplementary materials.

Note that the $n = 81$ *experimental* comparisons (by definition) compare PS-I with I-PS, where the PS and I phases are the same regardless of the sequence (PS-I or I-PS) in which they were carried out. Consequently, high (or, low) fidelity PS-I implementations are automatically compared with high (or, low) fidelity I-PS implementations. For the I-PS condition in the remaining $n = 85$ *quasi-experimental* comparisons, (a) 14.11% comparisons implemented the I phase using compare and contrast style of discussion, (b) 15.29% comparisons involved variants of scaffolding, sensemaking prompts, or additional learning resources

TABLE 2

Examples of PS-I designs with varying design fidelity to PF

Study characteristics	Loehr et al. (2014) (Study 1)	Loibl and Rummel (2014a) (Study 1)	Song (2018)
Problems affording multiple RSMs	Yes	Yes	Yes
Evidence for multiple RSM generation	No	No	Yes
Affective draw of the problem	No	Yes	Yes
Group work as the participation structure	No	Yes	Yes
Instruction building on student solutions	No	Yes	Yes
Social surround facilitation (problem-solving phase)	—	Low	High
Social surround facilitation (Instruction phase)	Monologue-dominant	Dialogue-dominant	Dialogue-dominant
Overall PF fidelity score (computed as a percentage)	16.66%	71.42%	100%
Hedge's *g* (effect size)	−0.97	1.31	0.73
95% Confidence interval	[−2.1, 0.16]	[0.22, 2.4]	[−0.31, 1.78]
Grade level	2nd to 5th graders	6th to 10th graders	6th to 10th graders
Geographical distribution	North America	Europe	Asia
Nature of pretesting	Targeted	Prerequisite	Targeted
Intervention type	Experimental	Quasi-experimental	Quasi-experimental
Intervention time span	A few hours	A few hours	A few days
Learning concept	Math	Math	Biology
Learning outcome	Conceptual knowledge	Conceptual knowledge	Clubbed outcomes

Note. PS-I = problem solving followed by instruction; PF = productive failure.

(e.g., in the form of worked examples) in the PS phase, and (c) 64.7% comparisons implemented group work as the participation structure in the PS phase. Note that despite the included comparisons using similar learning materials for the PS-I and I-PS conditions, we could not code for evidence for multiple RSM generation and affective draw in the PS phase of the I-PS implementations. This is because in the majority of the cases (e.g., Kapur & Bielaczyc, 2012; Loibl & Rummel, 2014a; Mazziotti et al., 2015; Schwartz & Martin, 2004), the PS phase following instruction was practice-oriented (rather than being generative in nature). In comparisons where students in the I-PS condition were provided the same invention problem as PS-I post-instruction (e.g., Chase & Klahr, 2017; Jarosz et al., 2017;

Kapur, 2012), they just used the taught canonical formulation to compute the correct answer right away. The fact that instruction limits spontaneous exploration and discovery is not unsurprising (Bonawitz et al., 2011). Additionally, several articles involving quasi-experimental comparisons (e.g., Fukaya et al., 2018; Hofer et al., 2018) stated that the I phase of I-PS was designed to provide opportunities for reflection and/or group discussion, that is, students were asked to engage with the presented worked examples, ask questions, and so on. However, none of these articles provided evidence that this was indeed the case, and therefore it is hard to speculate whether the I-PS conditions indeed used dialogue-dominant discourse style to facilitate the I phase.

Regression Analysis

To assess the combined effect of different PF fidelity criteria on the standardized effect sizes within our three-level meta-analytic model, we

1. Used a binary coding scheme, to give each criterion 1 point if it was present in the PF (or PS-I) implementation within a comparison (*high/low* surround facilitation during problem-solving was coded as 1/0; *dialogue-dominant/monologue-dominant* social surround facilitation during instruction was coded as 1/0).
2. Summed up these points to compute a synthesized PF fidelity score (*min* 0, *max* 7 points).
3. Divided the synthesized PF fidelity score by the total number of points achievable to compute a percentage score (remember that not all studies provided information about social surround facilitation during the problem-solving phase). This aggregated descriptive statistic for the raw data ($n = 166$ comparisons) highlighted the differential extent to which PS-I approaches incorporate elements of PF, and was distributed as follows: 0% to 25% ($n = 22$), 25% to 50% ($n = 60$), 50% to 75% ($n = 36$), 75% to 100% ($n = 48$).
4. Ran a meta-regression to estimate the predictive influence of this percentage score on the observed effect sizes. This gave us an account of if/whether and how much of variance might PF fidelity explain in the observed effect size estimates.

To unpack the relative importance of each of the seven PF fidelity criteria for prediction of effect size estimates, we further built a multiple-regression model. The seven binary-coded PF fidelity criteria served as independent variables,[8] and the standardized effect size was used as the dependent variable. Empirically, we used two approaches to select and insert predictors (here, the seven PF fidelity criteria) into this model. First, a *stepwise* model-building approach implemented using WEKA, a popular open-source machine learning toolkit (Witten et al., 2016) was used. Here, we successively inserted (forward selection) or deleted (backward selection) predictors in/from the multiple-regression model based on the amount of variability explained. Multicollinear predictors were removed in the final model creation. The predictors (PF fidelity criteria) were rank-ordered using Pearson's correlation between them and the dependent variable

(standardized effect size). In selecting the best predictor subset, the individual predictive ability of each predictor along with the degree of redundancy between them was considered, with a preference for subsets of predictors that were highly correlated with the dependent variable and had low intercorrelation. This evaluation produced a numeric measure (merit) of the expected performance of a subset.[9]

Second, a *multimodel inference* model-building approach implemented using R (Harrer et al., 2019), and frequently used in the meta-analysis literature was used. Here, we simultaneously assessed the evidential support for different multiple-regression models involving all possible combinations of predictors. Model fit was evaluated using corrected Akaike's Information Criterion (AICc; a statistical criterion that rewards simpler, more parsimonious models to avoid overfitting). We then synthesized estimated coefficients of predictors across all possible models to infer the importance of each PF fidelity criteria.[10]

Publication Bias

As part of the meta-analysis, we also investigated publication bias in the selected articles based on two different theoretical assumptions. First, we visualized whether small studies with small effect sizes were missing through funnel plots. The underlying assumption was that small studies are at greatest risk for being nonsignificant (and thus being missing), while large studies are likely to get published irrespective of the significance of results (e.g., due to large commitment of resources involved). Only small studies with a very large effect size become significant and are likely to be published. Thus, in the presence of publication bias, the funnel should look asymmetrical because only small studies with a large effect size would be published, while small studies without a significant, large effect would be missing. Additionally, we also used inferential statistics (Egger's test) for evaluating presence of asymmetry in the funnel plot and trim-and-fill procedures (Duval & Tweedie, 2000) if applicable.

Second, we explored an alternative way to assess the possibility of publication bias in our data. We visualized *p*-curves (distribution of significant *p*-values) for all included comparisons to examine whether our data reported more significant *p* values that were low ($p < .01$) rather than high ($.04 < p < .05$). True effects, those that differ from zero (e.g., more .01s than .04s), lead to right-skewed *p*-curves, and nonexistent effects lead to flat *p*-curves (as many .01s as .04s) or an equal probability of different significance levels (Simonsohn et al., 2015). We conducted follow-up statistical tests to assess whether the *p*-curve was significantly right-skewed (indicative of a true effect behind the data) and whether it was flat (indicative of insufficient power, or there being no true effect behind the data). Based on a significant right-skewness test and a nonsignificant flatness test, the presence or absence of evidential value in the *p*-curve can be ascertained (see Simonsohn et al., 2015, for details).[11]

Results

Pooled Effect Size Estimates

Results for Research Question 1 suggested that overall, the SMD or effect size for conceptual knowledge and transfer was moderate (Hedge's $g = 0.36$,

TABLE 3

Subgroup analysis focusing on PF fidelity

PF fidelity criteria	Hedge's *g* [95% *CI*]
Problems affording multiple RSMs [n.s.][a]	Yes (0.37 [−0.41, 1.38]) No (−0.11 [−0.99, 0.77])
Evidence for multiple RSM generation [+]	Yes (0.47 [0.00, 0.61]) No (0.16 [−0.08, 0.40])
Affective draw of the problem [n.s.]	Yes (0.44 [−0.08, 0.58]) No (0.19 [−0.08, 0.46])
Group work as the participation structure [*]	Yes (0.49 [0.01, 0.58]) No (0.19 [−0.02, 0.40])
Instruction building on student solutions [*]	Yes (0.56 [0.07, 0.64]) No (0.20 [0.01, 0.40])
Social surround facilitation [n.s.] (problem-solving phase)	High (0.58 [−0.30, 0.74]) Low (0.36 [0.02, 0.71])
Social surround facilitation [+] (instruction phase)	Dialogue-dominant (0.55 [0.00, 0.63]) Monologue-dominant (0.24 [0.05, 0.43])

Note. PF = productive failure; CI = confidence interval; RSM = representation and solution method; PS-I = problem solving followed by instruction; I-PS = instruction followed by problem solving. Positive effect sizes depict results in favor of PS-I (treated as the experimental condition). Negative effect sizes depict results in favor of I-PS (treated as the control/comparison condition).
[a]Significant subgroup differences are marked ($p < .1$: +, $p < .05$: *, $p < .01$: **, $p < .001$: ***, n.s.: not significant).

$p < .0001$, 95% *CI* [0.20, 0.51]), and in favor of PS-I.[12] The heterogeneity between comparisons was moderate ($Q[df = 165] = 295.91$, $p < .0001$). A total of 57.99%, 38.51%, and 3.50% of the variance were explained by the three levels of our meta-analytic model (total $I^2 = 42.01\%$). This three-level model (AICc 349.80, $p < .0001$) captured the variability in our data significantly better than a two-level model (AICc 370.83) that did not account for the nesting of studies within the articles. For the $n = 51/166$ comparisons assessing procedural knowledge, results suggested the pooled effect size for procedural knowledge to be Hedge's $g = -0.03$, $p = .7384$, 95% *CI* [−0.20, 0.15]. See the online supplementary materials (Table S1) for the forest plot, the standardized mean difference (SMD) for each comparison, 95% confidence intervals (lower CI, upper CI), and the contribution of each comparison in the pooled effect size calculation (%).

Subgroup Analysis

Subgroups accounting for heterogeneity in the overall effect size are shown in Tables 3, 4, and 5. Results for Research Question 2 suggested several significant subgroup differences as elaborated below.

Productive Failure Fidelity for the PS-I Design

Descriptively, comparisons had a higher effect size (in favor of PS-I), if any of the seven PF fidelity criteria were followed in the PS-I design. Only *four* of these

TABLE 4

Subgroup analysis focusing on students' incoming characteristics

Incoming student profile	Hedge's *g* [95% *CI*]
Age range (grade level) [**, +, n.s., *][a] (Others: postgraduates and professionals)	2nd to 5th graders (−0.09 [−0.92, −0.16])
	6th to 10th graders (0.50 [−0.04, 0.58])
	Undergraduates (0.28 [−0.46, 0.24])
	Others (1.03 [0.05, 1.39])
Geographical distribution [n.s., n.s., *, n.s.]	Europe (0.19 [−0.59, 0.15])
	North America (0.24 [−0.54, 0.09])
	Asia (0.64 [0.03, 0.73])
	Australia (0.95 [−0.13, 1.39])
Nature of pretesting [n.s., n.s., n.s.]	None (0.31 [−0.44, 0.30])
	Prerequisite (0.30 [−0.63, 0.50])
	Targeted (0.39 [−0.26, 0.42])

Note. PF = productive failure; CI = confidence interval; PS-I = problem solving followed by instruction; I-PS = instruction followed by problem solving. Positive effect sizes depict results in favor of PS-I (treated as the experimental condition). Negative effect sizes depict results in favor of I-PS (treated as the control/comparison condition).

[a]Significant subgroup differences are marked ($p < .1$: +, $p < .05$: *, $p < .01$: **, $p < .001$: ***, n.s.: not significant). For categories with >2 subgroups, significance refers to comparison of a particular subcategory with the rest of the subcategories.

seven subgroup differences were significant (or marginally significant)—evidence for multiple RSM generation ($p = .05$), group work participation structure ($p = .04$), instruction building on student solutions ($p = .02$), and social surround facilitation during the instruction phase ($p = .05$). The *overall PF fidelity score* (computed as a percentage) was a significant predictor of the effect size ($\beta = 0.0065$, $p < .001$, 95% *CI* [0.0044, 0.0087]).

We found that longer interventions had a higher overall PF fidelity score. A Welch two-sample *t* test ($t[154.22] = 8.504$, $p < .001$) revealed significant differences in overall PF fidelity score, with the mean of interventions spanning *a few days* being 74.45 ($SD = 22.64$) and those spanning *a few hours* being 42.01 ($SD = 26.13$). Significant chi-square tests for individual PF fidelity criteria also suggested that longer interventions were more likely to comprise these criteria. These observations led us to explore the potential *confound* of the length of the intervention (instead of the PF fidelity) being responsible for explaining the observed effects. Different meta-regression models were run to explicitly incorporate the interaction of intervention time span with (a) the overall PF fidelity score and (b) the presence of each of the seven individual PF fidelity criteria. In summary, we found that

1. Despite an overall increase in effect size with increase in overall PF fidelity ($\beta = 0.0173$, $p = .0005$, 95% *CI* [0.0077, 0.0269]), there was in fact a decrease for interventions spanning *a few hours* ($\beta = -0.0116$, $p = .0546$, 95% *CI* [−0.0235, 0.0002]). This suggests that cramming too many design features within a short amount of time may not be optimal.

TABLE 5

Subgroup analysis focusing on intervention characteristics

Intervention	Hedge's *g* [95% CI]
Type [n.s.][a]	Quasi-experimental (0.46 [−0.10, 0.50])
	Experimental (0.25 [0.04, 0.47])
Time span [n.s.]	A few days (0.41 [−0.22, 0.42])
	A few hours (0.32 [0.11, 0.52])
Learning concepts targeted [n.s., n.s., n.s., n.s., n.s., n.s., n.s., *]	Math (0.48 [−0.10, 0.52])
	Physics (0.39 [−0.34, 0.43])
	Chemistry (0.48 [−0.71, 0.96])
	Biology (0.32 [−0.73, 0.66])
	Medicine (0.24 [−0.60, 0.33])
	Psychology (1.38 [−0.43, 2.49])
	Environmental science (0.56 [−0.80, 1.20])
	Domain-general skills (−0.17 [−1.11, −0.02])
Learning outcomes assessed [n.s., n.s., n.s.]	Conceptual (0.33 [−0.30, 0.22])
	Transfer (0.40 [−0.19, 0.32])
	Clubbed (0.31 [−0.43, 0.31])

Note. PF = productive failure; CI = confidence interval; PS-I = problem solving followed by instruction; I-PS = instruction followed by problem solving. Positive effect sizes depict results in favor of PS-I (treated as the experimental condition). Negative effect sizes depict results in favor of I-PS (treated as the control/comparison condition).

[a]Significant subgroup differences are marked ($p < .1$: +, $p < .05$: *, $p < .01$: **, $p < .001$: ***, n.s.: not significant). For categories with >2 subgroups, significance refers to comparison of a particular subcategory with the rest of the subcategories.

2. Despite an overall increase in effect size for interventions comprising the *affective draw PF fidelity criteria* (relative to interventions where the affective draw of the problem was not considered—$\beta = 0.3710$, $p = .0056$, 95% *CI* [0.1101, 0.6319]), the increase was stronger for interventions spanning *a few hours* ($\beta = 0.7322$, $p = .0732$, 95% *CI* [−0.0695, 1.5339]). This suggests that for shorter interventions, affective draw of the problem may be critical.
3. Despite an overall increase in effect size for interventions comprising the *group work PF fidelity criteria* (relative to interventions where individual work was used as the participation structure—$\beta = 0.6286$, $p < .0001$, 95% *CI* [0.3628, 0.8945]), the increase was stronger when group work was *not implemented* in interventions spanning *a few hours* ($\beta = 0.7395$, $p = .0023$, 95% *CI* [0.1067, 1.3723]). This suggests that for shorter interventions, designing the problem-solving phase to accommodate individual work yields a better predictive estimate of the observed effect sizes.
4. Interaction effects of intervention time span with the remaining five PF fidelity criteria were not significant, lending *weak overall support* to the conjecture that intervention length (and not PF fidelity) might be responsible for the observed overall effects.

Students' Incoming Characteristics

Age range, as assessed by students' grade level, had (marginally) significant subgroup differences for all student subcategories (except for undergraduates). The pooled effect size estimate for younger students (2nd to 5th graders) was negative, and these estimates increased (or became more positive) with the age range. Furthermore, pretesting focused on learning concepts targeted during the intervention was descriptively associated with higher effect sizes (in favor of PS-I) relative to pretesting focusing on prerequisite concepts and cases of no pretesting. These subgroup differences were however not significant. This result, which reflects a facilitatory (rather than inhibitory) effect of targeted pretesting, lends support to the forward testing effect conjecture (outlined earlier in the coding rationale subsection). To explore the *confound* of age and the nature of pretesting,[13] we ran a meta-regression model by explicitly including their interaction as a predictor of the observed effect sizes. The interaction term was not significant (*p*s .15 to .77) in the model.

Another observation was the significant difference in the PF fidelity score that existed between 6th–10th graders and undergraduates (assessed using an ANOVA [$F(3, 162) = 6.485$, $p = .0004$] and follow-up Tukey HSD [$p = .0002$] pairwise comparisons). To explore whether the increasing trend of effect sizes in favor of PS-I across students' grade levels could therefore be explained by differential PF fidelity across these PS-I implementations, we ran another meta-regression model by explicitly including the *interaction of grade level and overall PF fidelity score* as a predictor of the observed effect sizes. The interaction term was (marginally) significant for two student subgroups—higher PF fidelity was associated with higher effect sizes in the case of 6th to 10th graders ($\beta = 0.0190$, $p < .0001$, 95% *CI* [0.0106, 0.0274]) and undergraduates ($\beta = 0.0080$, $p = .0934$, 95% *CI* [−0.0014, 0.0173]).

Finally, when looking at the distribution of effect sizes by *geographical regions*, interventions carried out in Australia and Asia had descriptively higher effect sizes than those conducted in Europe and North America. The only significant subgroup difference we found was for Asia (see Table 4), relative to the remaining continents.

Intervention Characteristics

When looking at intervention characteristics, effect sizes were descriptively higher (and in favor of PS-I) for (a) quasi-experimental comparisons relative to experimental comparisons and (b) interventions spanning a few days relative to those that spanned a few hours. These subgroup differences, were, however, not significant. We investigated two reasons why experimental comparisons might have relatively lower effect sizes.

1. Our *first* hypothesis was that lower manipulation flexibility in experimental studies might result in them having low overall PF fidelity, which could contribute to lower effect size. Indeed, a Welch two-sample *t*-test ($t[163.19] = -14.83$, $p < .001$) revealed significant differences in overall PF fidelity score (computed as a percentage), with the mean of *experimental* comparisons being 32.36 ($SD = 19.50$) and those of *quasi-experimental*

comparisons being 76.78 ($SD = 19.07$). However, explicitly incorporating the interaction of intervention type with PF fidelity score into a meta-regression did not reveal a significant interaction effect ($p = 0.79$).

2. Our *second* hypothesis was that experimental comparisons might be constrained by time, and a shorter duration (consequently, lesser time on task for students) might be responsible for relatively lower effect sizes. This possibility was empirically tested by conducting a χ^2 test for the presence of a relationship between intervention type and intervention time span, which turned out to be significant ($\chi^2(1) = 33.15$, $p < .001$). Follow-up pairwise nominal independence tests revealed that there were significantly more experimental comparisons that spanned *a few hours*, and significantly more quasi-experimental comparisons that spanned *a few days*. When explicitly modeling the interaction of intervention type and intervention time span via a meta-regression, we found that the predictive estimate for quasi-experimental comparisons spanning *a few hours* was in fact negative ($\beta = -0.5530$, $p = .0838$, 95% *CI* [−1.1807, 0.0748]). This suggests that the extent to which a quasi-experimental intervention reports high effect sizes favoring PS-I might be influenced by the length of the intervention.

With respect to *learning concepts* targeted in the interventions, moderate effect sizes in favor of PS-I were observed for domain-specific subjects (e.g., math, basic sciences, medicine). For learning domain-general skills, however, the effect size was low, and in favor of I-PS. Finally, separation of effects by *learning outcomes* suggested moderate effects for transfer (Hedge's $g = 0.40$, 95% *CI* [−0.19, 0.32]), and low to moderate effects for conceptual knowledge (Hedge's $g = 0.33$, 95% *CI* [−0.30, 0.22]) and clubbed assessments (Hedge's $g = 0.31$, 95% *CI* [−0.43, 0.31]). There were, however, no significant subgroup differences.

Ranking of Productive Failure Fidelity Criteria

Table 6 outlines result from multiple-regression analysis based on the two model building methods that we used to rank the seven PF fidelity criteria in terms of their importance to predicting effect size estimates. Our first model building approach (*stepwise regression*) revealed instruction building on student solutions, group work as the participation structure in the problem-solving phase, evidence for multiple RSM generation in the article, and dialogue-dominant social surround facilitation in the instruction phase as the *top four* most important predictors. Remember that for stepwise model-building methods, predictor ranking (importance) was based on Pearson's correlation between the predictor and the dependent variable. Furthermore, the latter three predictors ($\beta = 0.28$, $\beta = 0.27$, and $\beta = 0.20$, respectively) were also selected into the regression model via both forward and backward selection approaches. The merit of this subset of predictors was 0.358 (model $R^2 = 13.51\%$).

Our second model-building approach (*multimodel inference*) revealed evidence for multiple RSM generation in the article, group work, social surround facilitation in the instruction phase, and affective draw of the problem as the *top*

TABLE 6

Multiple-regression analysis to examine the relative importance of PF fidelity criteria

PF fidelity criteria	Stepwise model-building method		Multimodel inference model-building method	
	Predictor importance (Pearson correlation)	Predictor estimate (regression coefficient)	Predictor importance (Akaike weights)	Predictor estimate (regression coefficient)
Problems affording multiple RSMs (1 = yes, 0 = no)	0.17	0.12	0.38	0.03
Evidence for multiple RSM generation (1 = yes, 0 = no)	**0.25**	0.27	**0.93**	0.27
Affective draw of the problem (1 = yes, 0 = no)	0.18	—	**0.56**	0.09
Group work as the participation structure (1 = yes, 0 = no)	**0.29**	0.28	**0.85**	0.24
Instruction building on student solutions (1 = yes, 0 = no)	**0.28**	—	0.40	0.01
Social surround facilitation in PS phase (1 = high, 0 = low)	0.09	0.14	0.43	0.03
Social surround facilitation in I phase (1 = dialogue, 0 = monologue-dominant)	**0.23**	0.20	**0.60**	0.11

Note. RSM = representations and solution method; PS = problem solving.

four most important predictors. Remember that for multimodel inference, predictor ranking (importance) was based on the sum of Akaike weights for each predictor in the subset of regression models it appeared. Taken together, these rankings align with those from the stepwise model-building method.

Publication Bias

Small-Study Bias

The funnel plot visualization to detect if small studies with small effect sizes were missing is shown in the left half of Figure 3. The absence of asymmetry is visually evident. Egger's test of asymmetry was not significant (*intercept* = 0.63, $p = .219$). Since this source of publication bias could be therefore ruled out, no studies were imputed in trim and fill analysis.

P-*Curve Analysis*

The *p*-curve analysis as shown in the right half of Figure 3, however, revealed interesting results. Visually, the *p*-curve looked skewed to the right. The estimated power of the comparisons included in the analysis, which is useful both in designing future studies and in interpreting existing results (Gelman & Carlin,

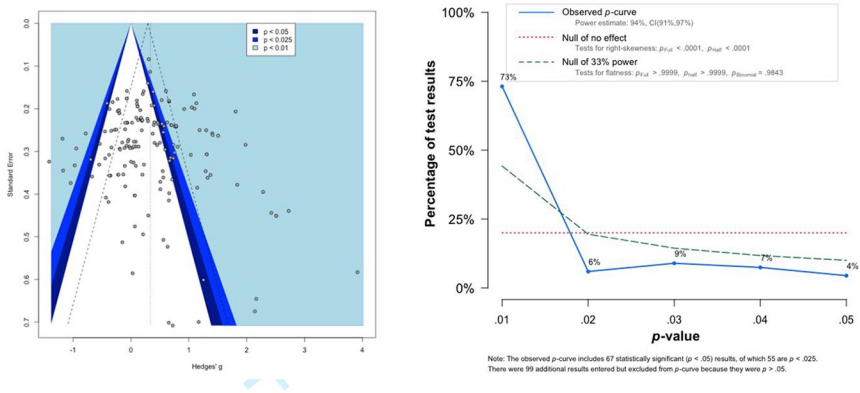

FIGURE 3. *Funnel plot (left) and p-curve (right) to assess publication bias for n = 166 PS-I and I-PS comparisons are depicted. Funnel plots are colored by the significance level into which the effect size of each of the comparisons fall. Bubbles in the plot depict included comparisons. The Y-axis shows the standard error (SE) of each comparison, with larger comparisons (which thus have a smaller SE) plotted on top of the Y-axis. The X-axis shows the standardized effect size of each comparison. Striped vertical line corresponds to the pooled effect size estimate. Color coding—dark blue (dark gray): p < .05, blue (gray): p < .025, light blue (light gray): p < .01. The p-curve is shown on the right with a blue/solid line. The Y-axis shows the percentage of comparisons. The X-axis shows the p-values.*

2014), was high (94%, *CI* [91%, 97%]). Numerically, the test for right-skewness was significant (meaning the *p*-curve was heavily right-skewed), and the test for flatness was not significant. Based on the results of the right-skewness and the flatness test, we could conclude that evidential value was present in the *p*-curve. The estimate of true effect size (in absence of publication bias) was therefore computed to be Hedge's $g = 0.87$. With a moderate (<50%) level of heterogeneity in included comparisons ($I^2 = 42.01\%$), we expect these estimates to be robust (van Aert et al., 2016).

Discussion

Overall, our results showed a significant effect (Hedge's *g*) of 0.36 [95% *CI* 0.20, 0.51] in favor of PS-I (compared to I-PS) for conceptual knowledge and transfer, and a nonsignificant effect (Hedge's *g*) of −0.03, 95% *CI* [−0.20, 0.15] for procedural knowledge. For procedural knowledge, the results align with prior discussions in the PS-I literature (Chen & Kalyuga, 2020; Loibl et al., 2017) suggesting that despite not being better than I-PS, PS-I to say the least, does not hurt or compromise on students' knowledge of procedures. It should be plausible to expect fluency in procedures over time with sufficient number of practice opportunities, something that is hard to capture within the relatively short time span of educational interventions. For conceptual knowledge and transfer, although the magnitude of the overall effect size might be moderate when based on Cohen's

benchmarks (0.2 small, 0.5 medium, 0.8 large), the *practical importance* of these estimates should be judged by the nature of the intervention being evaluated, its target population, and the outcome measure or measures being used (Hill et al., 2008). Kraft (2019), based on the distribution of 1942 effect sizes from 747 randomized control trials evaluating educational interventions with standardized test outcomes, has proposed a more plausible benchmark for interpreting effect sizes: <0.05 (small), 0.05 to 0.20 (medium), >0.20 (large). In that light, the impact of PS-I interventions evaluated via our meta-analysis can be considered rather large, relative to I-PS counterparts. We further found several significant subgroup differences that might explain efficacy of PS-I over I-PS. On accounting for evidence regarding publication bias in the included comparisons, estimation of true effect sizes for conceptual knowledge and/or transfer suggested a strong effect size (Hedge's *g*) in favor of PS-I (0.87). What might explain these patterns of results?

Productive Failure Fidelity for the PS-I Design

When compared to I-PS, convergent results from the subgroup and regression analysis suggested a strong trend of high PF design fidelity to be associated with higher effect sizes in favor of PS-I. According to Kapur and Bielaczyc (2012), PF embodies four core mechanisms (a) activation and differentiation of prior knowledge in relation to the targeted concepts, (b) attention to critical conceptual features, (c) explanation and elaboration of these features, and (d) organization and assembly of critical conceptual features into targeted concepts.

Mechanisms *a* and *b* are likely to be triggered by creating rich problems that engage students in design and use variant-invariant features to create opportunities for failure. By admitting multiple representations and solutions and offering intuitive hooks with an affective draw, such problems are well-poised to activate students' relevant prior knowledge and focus their attention on conceptual features of the problem. We operationalized these design features from Kapur and Bielaczyc (2012) into three concrete criteria: *problems affording multiple RSMs*, *evidence for multiple RSM generation*, and *affective draw of the problem*.

Mechanisms *b* and *c* are likely to be triggered by providing opportunities for explanation and elaboration via collaboration in mixed ability groups, support for students to collaborate through macro scripts, and pushing student thinking through disciplinary facilitation. Along with the appropriate participation structure, the social surround plays a key role during the initial problem-solving activity in creating a safe space to generate and explore ideas, setting and constantly emphasizing appropriate socio-mathematical norms and values, and providing affective support for persistence. We operationalized these design features from Kapur and Bielaczyc (2012) into two concrete criteria: *group work as the participation structure* and *social surround facilitation in the problem-solving phase*.

Mechanisms *b*, *c*, and *d* are likely to be triggered by providing opportunities to compare and contrast the affordances and constraints of failed or suboptimal RSMs and the assembly of canonical RSMs, which is achieved by having students explain their ideas and paraphrasing student explanations, comparing and contrasting these ideas to distill critical features, directing student attention to notice these critical features, and assembling the critical features into the canonical form.

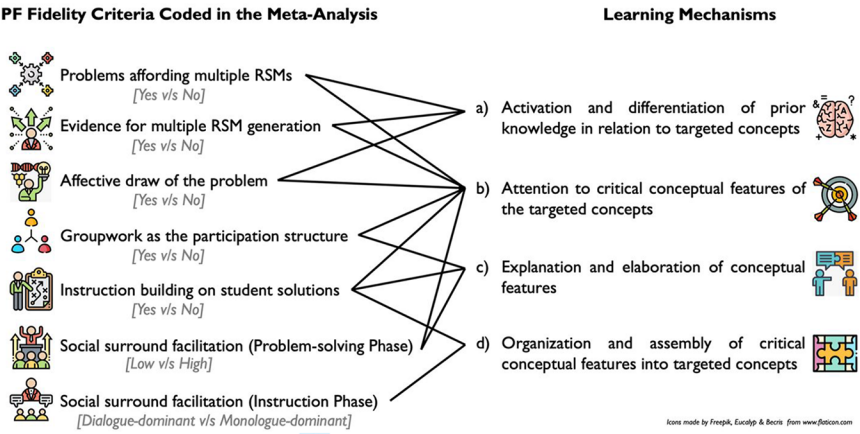

FIGURE 4. *Conjecture map between PF fidelity features and purported learning mechanisms.*

By focusing on the interrogation of students' ideas toward improvement and enhancing their participation during instruction by questioning, asking them to elaborate on each other's ideas, and so on, students can be provided opportunities to consolidate their knowledge gaps. We operationalized these design features from Kapur and Bielaczyc (2012) into two concrete criteria: *instruction building on student solutions* and *dialogue-dominant social surround facilitation in the instruction phase*.

By codifying PF design fidelity, we were able to systematically assess its impact on learning outcomes. The conjecture mapping between PF design fidelity criteria and the above-discussed learning mechanisms is summarized in Figure 4. It may not seem entirely surprising that PF works when its design fidelity is high (or, vice versa) when compared to I-PS. However, in the included comparisons, we often found instances of PS-I possessing low PF fidelity being compared with I-PS, and the resulting negative/null results being used to make claims about the nonefficacy of PF. Merely having the exact same sequence (problem solving followed by instruction) does not make PF and PS-I the same design. The extent to which results for PS-I can be used to make claims about PF depends on the extent to which PF fidelity criteria is followed in the PS-I learning design. Of course, PF design fidelity criteria is *not the only critical factor* that influences when PS-I performs better or worse than I-PS. As we have shown, there are several other important factors that might meaningfully categorize the data into less heterogeneous subgroups. But the question of fidelity is intrinsically fundamental. Similar to a manipulation check prior to reporting results of an intervention, it offered a useful starting point to investigate PF efficacy.

Finally, we found evidence for instruction building on student solutions, group work as the participation structure in the problem-solving phase, evidence for multiple RSM generation in the article, and dialogue-dominant social surround

facilitation in the instruction phase as being the *top four* most important PF fidelity predictors of effect sizes. These results from regression analysis offer a useful lens through which a designer may systematically go about choosing which PF fidelity criteria to implement during an inductive learning activity. Rather than forcing the implementation of all seven PF design fidelity criteria, it is important to consider what is feasible within the cultural context and the available time and resources.

When comparing PS-I with I-PS, it is also useful to critically reflect on whether the first stage of PF (i.e., the problem-solving phase focusing on prior knowledge activation and RSM generation) might in fact constitute a form of direct instruction. We argue that since the emphasis of this first stage is on *generating and interpreting solutions*, that is, students (typically *novices*) exploring solutions to complex problems based on concepts they have not formally studied, the initial problem-solving stage could be considered a form of discovery-based instruction, albeit one that serves to prepare students to learn from future (direct) instruction. Because no new canonical information is provided explicitly via cognitive scaffolds, worked examples, and so on, our view is that the first stage of PF *does not resemble* (direct) instruction, a learning experience where students are introduced to formal concepts of a domain in the most unambiguous (eloquent) way possible (Kirschner et al., 2006).

Students' Incoming Characteristics

We found that effects favoring PS-I over I-PS were higher if students belonged to a higher grade level. Younger age students (2nd to 5th graders), however, seemed to benefit from I-PS interventions, although the average effect size favoring I-PS was low (Hedge's g of −0.09 [95% CI −0.92, −0.16]. This is despite 72% (n = 18/25) of the comparisons involving 2nd to 5th graders providing evidence for high prior knowledge activation (reflected in the PF fidelity criteria of *evidence for multiple RSM generation*). However, the extent to which the activated prior knowledge is *relevant* for learning the targeted concepts can substantially affect students' performance. Research on goal specificity (e.g., Miller et al., 1999; Vollmeyer et al., 1996), mechanisms of errorful generation (Cyr & Anderson, 2015), as well as prior PS-I work (e.g., Kapur, 2015; Schwartz et al., 2011) has suggested that the benefits of prior knowledge activation such as noticing inconsistencies across multiple problem instances, encoding critical features from instruction, and so on, are contingent on the *relevance* of the activation. One explanation for the negative effect sizes disfavoring PS-I for young learners might therefore be the *low relevance* (albeit, *high activation*) of prior knowledge. Factors stemming from students' unawareness of productive metacognitive learning strategies (e.g., planning and monitoring solution generation efficacy, self-explaining, debugging, and error-evaluation) might alternatively contribute to this contrasting trend. Finally, across the comparisons involving 2nd to 5th graders, the low overall PF fidelity score (M = 56.86%, SD = 27.84%) might also be responsible for negative effects for PS-I. For these younger students then, scaffolding the initial problem-solving phase of PS-I to compensate for their insufficient prior knowledge about cognitive and metacognitive learning strategies might be critical. Richland and Simms (2015), more generally, have documented the importance of

scaffolding exploratory problem solving through a series of studies on induction within (non-) STEM domains. They emphasize explicit support in noticing the relevance of relational thinking, providing adequate processing resources to mentally hold and manipulate relations, and facilitating recognition of both similarities and differences when drawing analogies.

We further found that comparisons where students were pretested on concepts targeted in the intervention showed a higher effect size favoring PS-I (Hedge's *g* 0.39 [95% *CI* −0.26, 0.42]), relative to interventions where no pretest was held (Hedge's *g* 0.31 [95% *CI* −0.44, 0.30]), or the pretest targeted only prerequisite concepts (Hedge's *g* 0.30 [95% *CI* −0.63, 0.50]). Therefore, including a pretest targeting similar learning materials as the PS-I intervention itself may serve to enhance the differential advantages over I-PS (see Yang et al., 2018, for the facilitatory effects of forward testing), instead of diminishing PS-I's effectiveness due to potential overlap with the triggered learning mechanisms (Newman & DeCaro, 2019).

Intervention Characteristics

We found studies carried out in Asia and Australia to have higher effects in favor of PS-I, relative to those carried out in Europe and North America. Also, longer quasi-experimental interventions had higher effect sizes in favor of PS-I over I-PS, compared to experimental comparisons that often spanned a shorter duration. Our analysis showed that this could in-part be because quasi-experimental interventions had higher overall PF fidelity (especially for 6th to 10th graders and undergraduates comprising the majority).

Furthermore, results suggested that except for *domain general-skills*, most STEM and non-STEM learning domains had moderate effect sizes in favor of PS-I. Why might this be the case? When learning domain-general skills, Chase and Klahr (2017) suggest that the problem-solving phase in and of itself is less likely to provide implicit feedback about what goals to adopt during the inquiry process (that strongly impacts learning). For instance, students' goals in pursuing inquiry might be scientific (finding out whether a variable impacts an outcome) or engineering oriented (guaranteeing some desired outcome). In such situations, aligning their goals to a scientific one takes precedence over the relative ordering of the instruction phase in which this might happen. In the absence of explicit feedback regarding what problem-solving actions are actually failures, PS-I can therefore be expected to perform worse. Students might not be in a position to use their awareness of knowledge gaps to consolidate information during the instruction phase (Matlen & Klahr, 2013). More empirical work is however needed to generalize these claims. Note that an I-PS advantage for domain-general skills does not mean that no alternative instructional models exist for teaching such skills. For instance, educational games comprising storylines with teachable moments tightly integrated into the gameplay experience are not naturally described as either I-PS (or PS-I) sequence.

Implications

Results from our meta-analytic review hold important implications for teaching and learning. Our results suggest that preparatory problem-solving approaches

with high design fidelity to PF might be a powerful way to design for long-term learning, especially for students' conceptual knowledge and the ability to transfer their knowledge to other domains. Based on these results, we recommend that classroom cultures prioritize enculturating students into a sensemaking disposition via preparatory problem-solving approaches. Even when such approaches lead students to generate failed or suboptimal solutions, students' relevant prior knowledge activation provides opportunities for teachers to show them limitations of this prior knowledge. By discussion of tradeoffs of different approaches to a problem, teacher's emphasis on negative knowledge (Gartmeier et al., 2008) can facilitate reflection and increase students' certainty and efficiency of future actions. Initial failures then might indeed serve as stepping stones for success.

Limitations and Future Work

Meta-analysis has been shown to be an effective technique in evaluating broad-scale pedagogical changes (Freeman et al., 2014). To the best of our knowledge, the current meta-analysis provides the most comprehensive integrative review of the relative efficacy of preparatory problem solving and instruction-first approaches to date. However, it would be a mistake to interpret the comprehensive nature of this work as providing authoritative conclusions about the pattern of effects. The identified list of influencing factors is important but not exhaustive. Some limitations and future work avenues are worth discussing.

First, nearly 75% of all included comparisons (see Table 1) targeted learning concepts of math and physics, which were, in turn, skewed with particular subtopics (e.g., variance). With future PS-I work targeting less frequent topics such as statistics process control, fair division/distribution, crypt-arithmetic, and so on, a topic-wise (instead of domain-wise) subgroup analysis might become plausible. The complexity of such learning materials (often proxied using ill-defined constructs such as element interactivity) might be a critical boundary condition, especially since a recent PS-I work (Ashman et al., 2020) suggests that advantages of PS-I over I-PS may diminish with increasing complexity. Clearly, more studies are also needed in other STEM domains as well as in non-STEM domains (e.g., psychology, arts, history) before we can generalize the effectiveness of PS-I, even with high PF fidelity. More studies with postgraduates and professionals need to be conducted before we can begin talking about whether PS-I treatments might positively impact such populations.

Second, included comparisons also revealed a ~3× higher number (diversity) of research groups carrying out empirical PS-I work across Europe and North America (relative to Asia and Australia). Although meta-analytic pooling accounts for the precision of the effect size estimates for different studies (and we expect the overall pooled estimate to be robust), the low effect sizes for geographical regions of Europe and North America (see Table 4) might alternatively stem from a multiplicity of theoretical lenses that these researchers bring to bear—lenses which shape the learning design and come with different philosophical flavors and different legacy of associated educational practices. The lower effectiveness of PS-I in Europe and North America might also point toward the importance of teacher training and/or their domain knowledge necessary to guide students' learning with the PS-I design. Although discourse style

and building on student-generated solutions are two teacher characteristics we could reliably code, credentials such as the teacher's intellectual ability, knowledge of subject-area, and classroom management skills, if reported in future studies, are very important to consider (Looney, 2011) when evaluating PS-I's effectiveness.

Third, for the 59.64% of included comparisons that provided evidence for multiple RSM generation during the problem-solving phase, very few reported actual failure (and success) rates. Although we expect students who have not formally learned a targeted concept to mostly generate wrong (or, suboptimal) solutions when presented with an ill-structured problem, multiple RSM generation is a rather weak proxy for failure. Consistent reporting of failure rates for the problem-solving phase in future work will afford directly testing students' experienced failure as a potential moderator of the observed effects.

Fourth, individual differences might also differentially benefit students' learning via preparation-first or instruction-first approaches. However, we did not have any information about them in the current meta-analysis. As more work in the field begins to consistently factor in heterogeneity in students' approach to failure-driven and success-driven learning (see, for instance, Sinha et al. [2021] and Sinha & Kapur [2021] that investigate posttest differences, accounting for students' prior knowledge, effort regulation, self-esteem, goal orientation, and attitude toward mistakes simultaneously), we would be in a better position to study alternative explanations for the preparatory benefits of problem solving and other sensemaking approaches. During consideration of when and for whom PS-I or I-PS designs work, educators should also be mindful of ways to support the learning of those students who may be disadvantaged in pre-knowledge or cognitive capacity. The current meta-analysis suggests a dearth of studies involving students with learning disabilities and/or learning difficulties. As yet, it is therefore still unknown whether PS-I can be used to improve learning for such student populations, especially when they are integrated within regular education classrooms. Future research could usefully investigate this. Finally, recent work (Haimovitz & Dweck, 2016) also suggests that students' individual differences in learning from failure and success might be strongly influenced (predicted) by failure-mindsets possessed by parents (view of failure as debilitating or enhancing). Future work might consider evaluating contributions from the home.

Fifth, we focused only on the learning outcomes of conceptual knowledge and transfer. However, if a key goal of preparatory problem solving is to make students aware of what they do not know, negative knowledge (what is not part of a concept, what procedure does not work, and why) should be an important outcome to measure after students have received instruction. Surprisingly, only one study in the meta-analysis (Loibl & Leuders, 2018) explicitly emphasized the assessment of negative knowledge and found differential benefits compared to conceptual knowledge outcomes. More studies would need to have separate assessments of negative knowledge (rather than including such questions as part of a conceptual knowledge or transfer posttest), before we can discern the robustness of results.

Sixth, it might be argued that different learning designs are not directly experimentally comparable because of differences in learning objectives. For instance, even though the only difference between PS-I and I-PS in a strict experimental comparison might be thought of as reversed ordering of the problem-solving and instruction phases, often, the goals of preparatory problem solving (e.g., prior knowledge activation, knowledge gap awareness, deep feature identification) are quite different from problem solving that follows instruction (e.g., fluency via application of taught procedures; Kalyuga & Singh, 2016).

A final limitation of our work lies in the empirical decision to dichotomize PF fidelity criteria and other student and intervention characteristics. For instance, the coding scheme for criteria such as affective draw and social surround facilitation comprises different aspects, and a finer-grained division into subcategories is plausible. However, in the process of iteratively developing these coding schemes, we realized that there was an insufficient number of comparisons where such subcategories could be coded with high enough interrater reliability. Having two subcategories simplifies interpretation, and aligns with the goals of establishing relationships between high-level PF fidelity criteria and effect sizes. Taken together, a consideration of these limitations warrants more transparent collection, assessment, and reporting of factors beyond those studied in this review. For future studies, this may even imply experimentally manipulating such factors to assess their *causal* impact.

Conclusion

Our meta-analysis brings robust and generalizable empirical evidence to bear on the long-standing debate about the pedagogical effectiveness of starting to teach a new concept with problem solving or instruction (Tobias & Duffy, 2009). By including an impressive number of comparisons from the burgeoning PF (and more generally PS-I) literature, we clearly established the efficacy of PS-I over I-PS when PF design fidelity was high, for the learning outcomes of conceptual knowledge and transfer. To the best of our knowledge, this is the first integrative empirical counterevidence against the claim that problem solving as a sensemaking activity for novices is not an effective instructional approach (cf. Kirschner et al., 2006, and follow-up work). By dividing the included comparisons into several meaningful subgroups based on students' grade level, intervention time span, and its (quasi-)experimental nature, we systematically highlighted why, when, for whom, and by how much might a PS-I design be superior to instruction-first approaches. After accounting for the significant evidence regarding publication bias, effects in favor of PS-I were even stronger. Our results advance the field by providing solid empirical evidence, in the context of existing literature, for using preparatory problem-solving approaches with high design fidelity to Productive Failure (Kapur & Bielaczyc, 2012), as a powerful way to design for long-term learning.

ORCID iD

Tanmay Sinha https://orcid.org/0000-0003-3069-2899

Notes

We thank Maya Spannagel (University of Zürich) for assistance with coding subgroup categories. We would also like to thank colleagues from the Professorship for Learning Sciences and Higher Education (ETH Zürich) for helpful feedback on previous iterations of the article.

[1] Google Scholar was chosen because it is currently the most comprehensive academic search engine with 389 million records (Gusenbauer, 2019). In addition, it allows access to both published and unpublished literature (e.g., dissertations).

[2] Since PF (Kapur, 2008) *pre-dates* PS-I (Loibl et al., 2017), we adopted more-inclusive search criteria of starting with citations to key PF articles. The three follow-up PF articles were published in flagship journals of the Cognitive Science Society (CSS), European Association for Research on Learning and Instruction (EARLI), and American Psychological Association (APA). The expected diversity in readership across these publication avenues led us to include citations from these articles in the search criteria, in addition to citations from the two seminal PF articles.

[3] For comparisons where information critical for pooling effect sizes was not present in the articles (*M*, *SD*, sample size), personal communication was established with relevant researchers between July 2019 and August 2019. We sincerely thank all researchers who promptly got back.

[4] Suppose one wants to teach students a math concept (and its associated procedures) that is novel to them, say standard deviation (*SD*). Procedural knowledge assessments would test application of the procedure for computing *SD* on a new dataset. Conceptual knowledge assessments would test the understanding of the critical features of *SD* and deducing its mathematical properties. Transfer assessments would test whether students can adapt knowledge of *SD* to solve problems on the concept of, say normalization, that is not explicitly covered in the instruction. Since nearly 90% of PS-I versus I-PS experimental comparisons assessed only conceptual knowledge or transfer outcomes (and not both), we did not use a *multivariate model*.

[5] Not all articles where the problem design afforded multiple RSMs necessarily comprised evidence of students actually generating multiple RSMs. Despite the lack of independence ($\chi^2(1,166) = 19.113,\ p < 0.001$), dependency metrics suggested a rather low level of association between these nominal variables (contingency coefficient = 0.321, $\varnothing$-coefficient = 0.339, Cramer's $V = 0.339$). Theoretically, since the quantity and quality of RSMs generated has been considered a proxy for the mechanism of *prior knowledge activation* in former PF literature (Kapur, 2014; Loibl & Rummel, 2014b), this is an important distinction. However, we acknowledge that if the diversity of the solution approaches was not part of the research question, authors of those included papers may not have provided information on the presence of multiple RSMs in their article (although students generated multiple solutions). Although our coding criteria accounts also for *qualitative* (and not just *quantitative*) descriptions of multiple RSM generation, running the risk of missing articles where students indeed generated multiple RSMs poses a threat to the coding validity.

[6] A total of 57.2% ($n = 95$) comparisons did not report any information on this variable; 33.8% of the remaining comparisons were annotated as *high*.

[7] A total of 1.8% ($n = 3$) comparisons did not report any information on this variable. As an illustrative example, the pretest for a PS-I intervention targeting the learning concept of standard deviation would be coded as *prerequisite* if the questions comprise only basic descriptive statistics (e.g., mean, median), and *targeted* if the questions additionally also comprise standard deviation.

[8] Since social surround facilitation during the problem-solving phase had 57.2% missing (unreported) data, we imputed these values based on the observed data distribution for

this criterion (by randomly drawing $n = 5$ subsets of the coded subcategories), prior to the multiple-regression analysis. Simply discarding missing data (and corresponding comparisons) may result in the complete cases being no longer representative of the target population, and consequently, estimates derived from them being subject to nonreporting bias. The predictor ranking and regression coefficients we report in Table 6 are therefore based on averaging the results from these five multiple-regression models that take as input a particular imputed version of social surround facilitation during the problem-solving phase (each time), along with rest of the PF fidelity criteria.

[9] Intuitively, the merit of a feature subset S containing k features is $\left[k * r_{cf}\right] / \sqrt{k + k(k-1) r_{ff}}$, where r_{cf} is the mean feature-class correlation, and r_{ff} is the average feature-feature intercorrelation. The numerator provides an indication of how predictive of the class a set of features are; the denominator of how much redundancy there is among the features (Hall, 1999).

[10] AICc estimates the relative Kullback–Leibler distance between a fitted candidate regression model and the data-generating mechanism. The relative weight of evidence (or Akaike weights) of a model i can be interpreted as the probability of model i being the best-approximating model from the entire set of candidate models. We use the sum of Akaike weights for each explanatory (independent) variable in the subset of regression models it appeared as a measure of its *importance* (Buckland et al., 1997). Values > 0.8 are considered high.

[11] With evidence for publication bias, the true effect size can be estimated. For an observed set of significant results, one can identify the expected *p*-curve that most closely resembles the observed *p*-curve, and then identify the effect size estimate corresponding to that *p*-curve. Because the shape of the *p*-curve is a function exclusively of sample size and effect size (and the sample size is observed), we can simply find the effect size that obtains the best overall fit (Harrer et al., 2019).

[12] Results from a complementary *Bayesian* multilevel meta-analysis model (Harrer et al., 2019; Williams et al., 2018) with *weakly informative priors* for true pooled effect size ($\mu \sim N(0, 1)$) and between-study heterogeneity ($\tau \sim HC(0, 0.5)$) aligned with these pooled effect size estimates: Hedge's $g = 0.36$, 95% *CI* [0.19, 0.52]. Based on recommendations in Harrer et al. (2019), the overall *Bayesian* methodology involved 4,000 iterations of the Markov Chain Monte Carlo–based NUTS sampling procedure (Hoffman & Gelman, 2014). Based on the empirical cumulative distribution function of the posterior distribution for the pooled effect size, the probability of the SMD being <0.2: 3.05%; <0.3: 25.11%; <0.4: 70.41%; <0.5: 95.72%.

[13] Previous reviews regarding the *forward testing* effect (Yang et al., 2018) suggest inconclusive evidence for the extent to which it might generalize to younger children (and older adults). Therefore, within the PS-I experimental paradigm, evaluating the interaction effects of age (grade-level) of students and whether a pretest was carried out might contribute toward understanding whether the forward testing effect is differentially pronounced across grade levels.

References

Ashman, G., Kalyuga, S., & Sweller, J. (2020). Problem-solving or explicit instruction: Which should go first when element interactivity is high? *Educational Psychology Review*, *32*(1), 229–247. https://doi.org/10.1007/s10648-019-09500-5

Assink, M., & Wibbelink, C. J. M. (2016). Fitting three-level meta-analytic models in R: A step-by-step tutorial. *Quantitative Methods for Psychology*, *12*(3), 154–174. https://doi.org/10.20982/tqmp.12.3.p154

Belenky, D. M., & Nokes-Malach, T. J. (2012). Motivation and transfer: The role of mastery-approach goals in preparation for future learning. *Journal of the Learning Sciences*, *21*(3), 399–432. https://doi.org/10.1080/10508406.2011.651232

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*(3), 322–330. https://doi.org/10.1016/j.cognition.2010.10.001

Buckland, S. T., Burnham, K. P., & Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics*, *53*(2), 603–618. https://doi.org/10.2307/2533961

Chase, C. C., & Klahr, D. (2017). Invention versus direct instruction: For some content, it's a tie. *Journal of Science Education and Technology*, *26*(6), 582–596. https://doi.org/10.1007/s10956-017-9700-6

Chen, O., & Kalyuga, S. (2020). Exploring factors influencing the effectiveness of explicit instruction first and problem-solving first approaches. *European Journal of Psychology of Education*, *35*(3), 607–624. https://doi.org/10.1007/s10212-019-00445-5

Chi, M. T. H. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in instructional psychology* (pp. 161–238). Lawrence Erlbaum.

Cobb, P. (1995). Cultural tools and mathematical learning: A case study. *Journal for Research in Mathematics Education*, *26*(4), 362–385. https://doi.org/10.2307/749480

Cyr, A. A., & Anderson, N. D. (2015). Mistakes as stepping stones: Effects of errors on episodic memory among younger and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 841–850. https://doi.org/10.1037/xlm0000073

Darabi, A., Arrington, T. L., & Sayilir, E. (2018). Learning from failure: A meta-analysis of the empirical studies. *Educational Technology Research and Development*, *66*(5), 1101–1118. https://doi.org/10.1007/s11423-018-9579-9

DeCaro, D. A., DeCaro, M. S., & Rittle-Johnson, B. (2015). Achievement motivation and knowledge development during exploratory learning. *Learning and Individual Differences*, *37*(January), 13–26. https://doi.org/10.1016/j.lindif.2014.10.015

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463. https://doi.org/10.1111/j.0006-341X.2000.00455.x

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(23), 8410–8415. https://doi.org/10.1073/pnas.1319030111

Fukaya, T., Uesaka, Y., & Ichikawa, S. (2019). Investigating the effects of thinking after instruction approach. *Educational Technology Research*, *41*(1), 1–11. https://doi.org/10.15077/etr.42105

Gartmeier, M., Bauer, J., Gruber, H., & Heid, H. (2008). Negative knowledge: Understanding professional learning and expertise. *Vocations and Learning*, *1*(2), 87–103. https://doi.org/10.1007/s12186-008-9006-1

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Glogger-Frey, I., Fleischer, C., Grüny, L., Kappich, J., & Renkl, A. (2015). Inventing a solution and studying a worked solution prepare differently for learning from direct instruction. *Learning and Instruction*, *39*(October), 72–87. https://doi.org/10.1016/j.learninstruc.2015.05.001

Gusenbauer, M. (2019). Google scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics*, *118*(1), 177–214. https://doi.org/10.1007/s11192-018-2958-5

Haimovitz, K., & Dweck, C. S. (2016). What predicts children's fixed and growth intelligence mind-sets? Not their parents' views of intelligence but their parents' views of failure. *Psychological Science*, *27*(6), 859–869. https://doi.org/10.1177/0956797616639727

Hall, M. A. (1999). *Correlation-based feature selection for machine learning* [Doctoral dissertation, The University of Waikato Hamilton]. https://www.cs.waikato.ac.nz/~mhall/thesis.pdf

Harrer, M., Cuijpers, P., Furukawa, T., & Ebert, D. D. (2019). *Doing meta-analysis in R* (Version 1.0.0) [Computer software]. https://doi.org/10.5281/zenodo.2551803

Hidi, S., & Harackiewicz, J. M. (2000). Motivating the academically unmotivated: A critical issue for the 21st century. *Review of Educational Research*, *70*(2), 151–179. https://doi.org/10.3102/00346543070002151

Higgins, J. P., & Green, S. (2011). *Cochrane handbook for systematic reviews of interventions*. John Wiley.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172–177. https://doi.org/10.1111/j.1750-8606.2008.00061.x

Hofer, S. I., Schumacher, R., Rubin, H., & Stern, E. (2018). Enhancing physics learning with cognitively activating instruction: A quasi-experimental classroom intervention study. *Journal of Educational Psychology*, *110*(8), 1175–1191. https://doi.org/10.1037/edu0000266

Hoffman, M. D., & Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623. https://doi.org/10.5555/2627435.2638586

Hsu, C.-Y., Kalyuga, S., & Sweller, J. (2015). When should guidance be presented in physics instruction? *Archives of Scientific Psychology*, *3*(1), 37–53. https://doi.org/10.1037/arc0000012

Jarosz, A. F., Goldenberg, O., & Wiley, J. (2017). Learning by invention: Small group discussion activities that support learning in statistics. *Discourse Processes*, *54*(4), 285–302. https://doi.org/10.1080/0163853X.2015.1129593

Kalyuga, S., & Singh, A.-M. (2016). Rethinking the boundaries of cognitive load theory in complex learning. *Educational Psychology Review*, *28*(4), 831–852. https://doi.org/10.1007/s10648-015-9352-0

Kapur, M. (2008). Productive failure. *Cognition and Instruction*, *26*(3), 379–424. https://doi.org/10.1080/07370000802212669

Kapur, M. (2012). Productive failure in learning the concept of variance. *Instructional Science*, *40*(4), 651–672. https://doi.org/10.1007/s11251-012-9209-6

Kapur, M. (2014). Productive failure in learning math. *Cognitive Science*, *38*(5), 1008–1022. https://doi.org/10.1111/cogs.12107

Kapur, M. (2015). The preparatory effects of problem solving versus problem posing on learning from instruction. *Learning and Instruction*, *39*(October), 23–31. https://doi.org/10.1016/j.learninstruc.2015.05.004

Kapur, M. (2016). Examining productive failure, productive success, unproductive failure, and unproductive success in learning. *Educational Psychologist*, *51*(2), 289–299. https://doi.org/10.1080/00461520.2016.1155457

Kapur, M., & Bielaczyc, K. (2012). Designing for productive failure. *Journal of the Learning Sciences*, *21*(1), 45–83. https://doi.org/10.1080/10508406.2011.591717

Kim, B., Pathak, S. A., Jacobson, M. J., Zhang, B., & Gobert, J. D. (2015). Cycles of exploration, reflection, and consolidation in model-based learning of genetics. *Journal of Science Education and Technology*, *24*(6), 789–802. https://doi.org/10.1007/s10956-015-9564-6

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*(2), 75–86. https://doi.org/10.1207/s15326985ep4102_1

Kraft, M. A. (2019). *Interpreting effect sizes of education interventions* (EdWorking Paper No. 19–10). Annenberg Institute for School Reform at Brown University. https://doi.org/10.26300/8pjp-2z74

Lamnina, M., & Chase, C. C. (2019). Developing a thirst for knowledge: How uncertainty in the classroom influences curiosity, affect, learning, and transfer. *Contemporary Educational Psychology*, *59*(October), 101785. https://doi.org/10.1016/j.cedpsych.2019.101785

Lazonder, A. W., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning: Effects of guidance. *Review of Educational Research*, *86*(3), 681–718. https://doi.org/10.3102/0034654315627366

Loehr, A. M., Fyfe, E. R., & Rittle-Johnson, B. (2014). Wait for it . . . Delaying instruction improves mathematics problem solving: A classroom study. *Journal of Problem Solving*, *7*(1), Article 5. https://doi.org/10.7771/1932-6246.1166

Loibl, K., & Leuders, T. (2018). Errors during exploration and consolidation: The effectiveness of productive failure as sequentially guided discovery learning. *Journal für Mathematik-Didaktik*, *39*(1), 69–96. https://doi.org/10.1007/s13138-018-0130-7

Loibl, K., & Leuders, T. (2019). How to make failure productive: Fostering learning from errors through elaboration prompts. *Learning and Instruction*, *62*(August), 1–10. https://doi.org/10.1016/j.learninstruc.2019.03.002

Loibl, K., Roll, I., & Rummel, N. (2017). Towards a theory of when and how problem solving followed by instruction supports learning. *Educational Psychology Review*, *29*(4), 693–715. https://doi.org/10.1007/s10648-016-9379-x

Loibl, K., & Rummel, N. (2014a). Knowing what you don't know makes failure productive. *Learning and Instruction*, *34*(December), 74–85. https://doi.org/10.1016/j.learninstruc.2014.08.004

Loibl, K., & Rummel, N. (2014b). The impact of guidance during problem-solving prior to instruction on students' inventions and learning outcomes. *Instructional Science*, *42*(3), 305–326. https://doi.org/10.1007/s11251-013-9282-5

Looney, J. (2011). Developing high-quality teachers: Teacher evaluation for improvement. *European Journal of Education*, *46*(4), 440–455. https://doi.org/10.1111/j.1465-3435.2011.01492.x

Matlen, B. J., & Klahr, D. (2013). Sequential effects of high and low instructional guidance on children's acquisition of experimentation skills: Is it all in the timing? *Instructional Science*, *41*(3), 621–634. https://doi.org/10.1007/s11251-012-9248-z

Mazziotti, C., Loibl, K., & Rummel, N. (2015). Collaborative or individual learning within productive failure: Does the social form of learning make a difference? In O. Lindwall, P. Häkkinen, T. Koschman, P. Tchounikine, & S. Ludvigsen (Eds.), *Exploring the material conditions of learning: The Computer Supported Collaborative Learning (CSCL) Conference 2015* (*Vol. 2*, pp. 570–575). International Society of the Learning Sciences.

Mazziotti, C., Rummel, N., Deiglmayr, A., & Loibl, K. (2019). Probing boundary conditions of productive failure and analyzing the role of young students' collaboration. *NPJ Science of Learning*, *4*(1), 1–9. https://doi.org/10.1038/s41539-019-0041-5

Miller, C. S., Lehman, J. F., & Koedinger, K. R. (1999). Goals and learning in microworlds. *Cognitive Science*, *23*(3), 305–336. https://doi.org/10.1207/s15516709cog2303_2

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine*, *151*(4), 264–269. https://doi.org/10.7326/0003-4819-151-4-200908180-00135

Newman, P. M., & DeCaro, M. S. (2019). Learning by exploring: How much guidance is optimal? *Learning and Instruction*, *62*(August), 49–63. https://doi.org/10.1016/j.learninstruc.2019.05.005

Nokes-Malach, T. J., Richey, J. E., & Gadgil, S. (2015). When is it better to learn together? Insights from research on collaborative learning. *Educational Psychology Review*, *27*(4), 645–656. https://doi.org/10.1007/s10648-015-9312-8

Richland, L. E., & Simms, N. (2015). Analogy, higher order thinking, and education. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(2), 177–192. https://doi.org/10.1002/wcs.1336

Schneider, B., & Blikstein, P. (2018). Tangible user interfaces and contrasting cases as a preparation for future learning. *Journal of Science Education and Technology*, *27*(4), 369–384. https://doi.org/10.1007/s10956-018-9730-8

Schneider, M., & Stern, E. (2010). The cognitive perspective on learning: Ten cornerstone findings. In H. Dumont, D. Istance, & F. Benavides (Eds.), *The nature of learning: Using research to inspire practice* (pp. 69–90). OECD. https://doi.org/10.1787/9789264086487-5-en

Schraw, G., Flowerday, T., & Lehman, S. (2001). Increasing situational interest in the classroom. *Educational Psychology Review*, *13*(3), 211–224. https://doi.org/10.1023/A:1016619705184

Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. *Cognition and Instruction*, *16*(4), 475–5223. https://doi.org/10.1207/s1532690xci1604_4

Schwartz, D. L., Chase, C. C., Oppezzo, M. A., & Chin, D. B. (2011). Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. *Journal of Educational Psychology*, *103*(4), 759–775. https://doi.org/10.1037/a0025140

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction*, *22*(2), 129–184. https://doi.org/10.1207/s1532690xci2202_1

Sears, D. A. (2006). *Effects of innovation versus efficiency tasks on collaboration and learning* [Doctoral dissertation, Stanford University]. https://iase-web.org/documents/dissertations/06.Sears.pdf

Sherin, B. L. (2000). Meta-representation: An introduction. *Journal of Mathematical Behavior*, *19*(4), 385–398. https://doi.org/10.1016/S0732-3123(01)00051-7

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better *p*-curves: Making *p*-curve analysis more robust to errors, fraud, and ambitious *p*-hacking, a reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General*, *144*(6), 1146–1152. https://doi.org/10.1037/xge0000104

Sinha, T., & Kapur, M. (2019). When productive failure fails. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *41*, 2811–2817.

Sinha, T., & Kapur, M. (2021). Robust effects of the efficacy of explicit failure-driven scaffolding in problem-solving prior to instruction: A replication and extension. *Learning and Instruction*, *75*(October), 101488. https://doi.org/10.1016/j.learninstruc.2021.101488

Sinha, T., Kapur, M., West, R., Catasta, M., Hauswirth, M., & Trninic, D. (2021). Differential benefits of explicit failure-driven and success-driven scaffolding in problem-solving prior to instruction. *Journal of Educational Psychology*, *113*(3), 530–555. https://doi.org/10.1037/edu0000483

Song, Y. (2018). Improving primary students' collaborative problem solving competency in project-based science learning with productive failure instructional design in a seamless learning environment. *Educational Technology Research and Development*, *66*(4), 979–1008. https://doi.org/10.1007/s11423-018-9600-3

Stockard, J., Wood, T. W., Coughlin, C., & Rasplica Khoury, C. (2018). The effectiveness of direct instruction curricula: A meta-analysis of a half century of research. *Review of Educational Research*, *88*(4), 479–507. https://doi.org/10.3102/0034654317751919

Thomas, D., & Brown, J. S. (2007). The play of imagination: Extending the literary mind. *Games and Culture*, *2*(2), 149–172. https://doi.org/10.1177/1555412007299458

Tobias, S., & Duffy, T. M. (2009). *Constructivist instruction: Success or failure?* Routledge. https://doi.org/10.4324/9780203878842

van Aert, R. C., Wicherts, J. M., & van Assen, M. A. (2016). Conducting meta-analyses based on *p* values: Reservations and recommendations for applying *p*-uniform and *p*-curve. *Perspectives on Psychological Science*, *11*(5), 713–729. https://doi.org/10.1177/1745691616650874

VanLehn, K. (1999). Rule-learning events in the acquisition of a complex skill: An evaluation of CASCADE. *Journal of the Learning Sciences*, *8*(1), 71–125. https://doi.org/10.1207/s15327809jls0801_3

Vollmeyer, R., Burns, B. D., & Holyoak, K. J. (1996). The impact of goal specificity on strategy use and the acquisition of problem structure. *Cognitive Science*, *20*(1), 75–100. https://doi.org/10.1207/s15516709cog2001_3

Williams, D. R., Rast, P., & Bürkner, P.-C. (2018). *Bayesian meta-analysis with weakly informative prior distributions*. PsyArXiv Preprints. https://doi.org/10.31234/osf.io/7tbrm

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Yang, C., Potts, R., & Shanks, D. R. (2018). Enhancing learning and retrieval of new information: A review of the forward testing effect. *NPJ Science of Learning*, *3*(1), Article 8. https://doi.org/10.1038/s41539-018-0024-y

Authors

TANMAY SINHA is a postdoctoral research fellow at ETH Zürich, Clausiusstrasse 59, Zurich 8092, Switzerland; email: *tanmay.sinha@gess.ethz.ch*. He completed his DrSc in learning sciences at ETH Zürich (2020), where his research focused on the pedagogical value of deliberate, guided failure in problem solving and the facilitatory effects of negative emotions in learning. Prior to joining ETH, he completed an MS in computer science at Carnegie Mellon University (2016), where his research focused on the computational modeling of multimodal human behavior, in particular the dynamics of curiosity in open-ended group work and the development of interpersonal closeness in dyadic peer tutoring. More generally, he is interested in the study of social and interpersonal factors in learning, and the design of educational interventions to scaffold failure-driven learning. To study learning-related phenomena, he has leveraged data sources from both online interaction (e.g., fine-grained clickstream, problem-solving activity, discussion forum behavior) and offline interaction (e.g., verbal, vocal, and visual communicative cues).

MANU KAPUR is a full professor at the Department of Humanities, Social and Political Sciences of the ETH Zürich (Clausiusstrasse 59, Zurich 8092, Switzerland; email: *manukapur@ethz.ch*) and holds the chair of Learning Sciences and Higher Education. Prior to joining ETH Zürich, he was a professor of psychological studies at The Education University of Hong Kong. He also worked as the head of the Curriculum, Teaching and Learning Academic Group as well as the head of Learning Sciences Lab at the National Institute of Education of Singapore. He conceptualized the notion of *productive failure* and has used it to explore the hidden efficacies in the seemingly failed efforts of small groups solving complex problems collaboratively in an online environment. He has done extensive work in real-field ecologies of mathematics classrooms to extend his work on productive failure across a range of schools in Singapore.