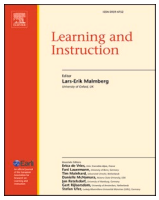
Learning and Instruction 75 (2021) 101488

Contents lists available at ScienceDirect

Learning and Instruction

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



Robust effects of the efficacy of explicit failure-driven scaffolding in problem-solving prior to instruction: A replication and extension

Tanmay Sinha *, Manu Kapur

ETH Zürich, Switzerland

ARTICLE INFO

Keywords:
Data science education
Higher education
Failure
Scaffolding
Problem-based learning

ABSTRACT

Although Productive Failure has shown to be effective (Kapur, 2016; Loibl, Roll, & Rummel, 2017), it is not clear if failure in problem-solving is necessary. Initial work in a quasi-experimental setting suggests that explicitly designing for experiences of failure leads to better learning outcomes than designing for success. We build on this to report on a controlled experimental study where students are exposed to failure-driven, success-driven, or no explicit scaffolding in problem-solving prior to instruction. For assessments of non-isomorphic conceptual understanding, our results align with those from prior work. Despite the similarity in posttest scores, students exposed to failure-driven scaffolding demonstrate higher quality of constructive reasoning than those receiving success-driven scaffolding. Additionally, our study reveals learning benefits of failure-driven scaffolding (for both posttest scores and reasoning quality) on assessments of transfer. Several cognitive, affective and meta-cognitive mechanisms are investigated to explain robust learning benefits of failure-driven scaffolding in preparatory problem-solving.

1. Introduction

Problem-solving followed by instruction (PS-I) comprises an initial generative problem-solving phase requiring students to generate multiple solutions to a problem. A formal instruction phase subsequently introduces targeted concepts and the canonical solution. PS-I learning designs activate students' prior knowledge, raise awareness of knowledge gaps, and aid in recognition of deep problem features (Loibl et al., 2017). These preparatory benefits have been posited to account for the relative efficacy of PS-I over instruction-first approaches, as evidenced by effect sizes favoring PS-I in a recent meta-analysis of the field (Hedge's g 0.36 [95% CI 0.20, 0.51], $N = 166$ comparisons) by Sinha & Kapur (2021a). Given such a robust trend of the superiority of PS-I designs across diverse learning domains and student populations, research into improving its effectiveness has intensified.

One such approach is to explicitly scaffold the initial problem-phase, typically towards success, by providing cognitive or metacognitive support (Holmes, Day, Park, Bonn, & Roll, 2014; Kapur, 2011; Loibl & Rummel, 2014a). However, meta-analytic evidence from Sinha & Kapur (2021b) suggests that relative to an unscaffolded PS-I design, such scaffolding attempts have largely been unsuccessful (Hedge's $g = -0.08$, 95% CI [-0.20, 0.04], $N = 60$ comparisons). An alternative approach to improve the effectiveness of preparatory problem-solving is explicit failure-driven scaffolding. To our knowledge, only one recent study has shown that nudging students towards suboptimal solutions via explicit failure-driven scaffolds may lead to stronger conceptual understanding than nudging students towards optimal solutions via explicit success-driven scaffolds (Sinha et al., 2020). However, this was a classroom-based study and not a fully controlled experiment.

Further, to the best of our knowledge, there is no research within PS-I comparing explicit failure-driven scaffolding to unscaffolded preparatory problem-solving. PS-I designs are usually aimed at introducing new concepts to novices in a domain. Therefore, they typically have high failure-rates even without any failure-scaffolding. This naturally leads to the question of whether there is an added efficacy of explicit scaffolding prior to instruction. Before emphasizing the design of explicit failure-driven experiences prior to formal instruction, thorough testing and replication in more controlled environments with similar student demographics and task domains is needed. That is precisely the aim of our study.

2. Theoretical background

We focus on differential preparatory effects of failure-driven and

* Corresponding author. Learning Sciences and Higher Education, ETH Zürich, RZ J 5, Clausiusstrasse 59, 8092, Switzerland.
E-mail addresses: tanmay.sinha@gess.ethz.ch (T. Sinha), manukapur@ethz.ch (M. Kapur).

https://doi.org/10.1016/j.learninstruc.2021.101488
Received 28 March 2020; Received in revised form 21 February 2021; Accepted 23 March 2021
Available online 10 May 2021

success-driven scaffolding within PS-I. To better situate our experimental replication context, we start by describing the role of scaffolding in PS-I (cf. section 2.1), cognitive and affective mechanisms that can be posited to be triggered in scaffolded PS-I (cf. section 2.2), and finally individual differences in student characteristics that may affect learning from scaffolded PS-I (cf. section 2.3). At a theoretical level, we are interested in the extent to which learning mechanisms are differentially triggered in scaffolded PS-I. At a more practical level, we are interested in the generalizability of the impact of failure and success-driven scaffolding to our study context (data science education) and targeted population (postgraduates).

2.1. Preparatory problem-solving and scaffolding

To combine strengths and mitigate drawbacks of engaging students in standalone problem-solving or lecture, prior research (Kapur & Bielaczyc, 2012; Loibl et al., 2017; Schwartz & Martin, 2004) has proposed the PS-I design that implicates a temporal ordering between these two learning activities. Scaffolding, or administering just-in-time support to allow students to make meaningful progress in a problem-solving task (Wood, Bruner, & Ross, 1976), is one way to facilitate learning through preparatory problem-solving without shrinking the problem-space upfront (Hmelo-Silver, Duncan, & Chinn, 2007; Quintana et al., 2004). There has been a longstanding debate on whether scaffolded problem-solving should lean towards being more success-driven or failure-driven (Kapur, 2016; Lee & Anderson, 2013).

Scaffolding towards success is likely to result in high performance, as one might expect when working with the suggested correct procedure. A straightforward implication is that non-promising solution pathways can be curtailed early, as students focus on improvement and making things work. However, it is also plausible that students' focus on high performance insufficiently challenges prior knowledge, bypasses inquiry points, and comes at the expense of depth of understanding of underlying domain principles (Kapur, 2016; Soderstrom & Bjork, 2015). Hidslugh performance does not guarantee that students are aware of the inconsistency between what solution approach works in a given context and the extent to which it generalizes to future contexts (Schwartz, Chase, & Bransford, 2012).

Scaffolding towards failure, on the other hand, is likely to result in an initial dip in performance, as one might expect when working with the suggested incorrect procedure. Repeated failures can also be penalizing, in that they can increase self-doubt and stability of future failure expectancies leading to the absence of control (Mikulincer, 1994). Experiencing failures might make students more susceptible to negative affective reactions (Tulis & Ainley, 2011) and increase stickiness to self-generated suboptimal solutions (Johnson & Seifert, 1994). However, numerous theoretical lenses also speak to the importance of failure in problem-solving (see Kapur (2016) and Metcalfe (2017) for reviews). One common thread tying these frameworks is the predication that failures initiate explanation and reflection processes to make sense of something that is not immediately apparent. By maximizing information gained from each problem-solving failure, the route to discovery can be made tractable. Echoing this philosophy, Hammer (2000, p. 58) has remarked that "wrong thinking should be seen as productive if it helps develop resources for right thinking later on". Negative knowledge (Gartmeier, Bauer, Gruber, & Heid, 2008), or the knowledge of what is not part of a concept and what procedure does not work and why, resulting from deliberately-designed failure experiences might enhance reflection quality. Even if students can activate partial knowledge of what does not work and why during preparatory problem-solving, it might still serve as a strong foundation to acquire robust knowledge when exposed to instruction later.

The incommensurability between initial training performance (as assessed during the preparatory problem-solving phase of PS-I) and delayed testing performance (as assessed via posttest following the instruction phase of PS-I), an argument advanced by Kapur (2016), is of prime interest for the current article. Although on one hand, "substantial learning could occur in the absence of any discernible changes in performance … more recent research … have demonstrated the converse to also be true – specifically, that changes in short-term performance often bear no relationship to long-term learning" (Soderstrom & Bjork, 2015, p. 193). By tapping into multiple performance measures throughout our PS-I design, we evaluate the differential impact of scaffolding towards success and failure on training and testing phases.

2.2. Mechanisms underlying preparatory problem-solving

To open possible explanatory bases for why certain PS-I designs work better than others, research syntheses by Kapur (2016) and Loibl et al. (2017) have consolidated a set of cognitive mechanisms. These mechanisms include, but are not limited to.

1. Intentionally activating relevant prior knowledge (*prior knowledge activation*)
2. Enhancing students' awareness of the problem situation and own knowledge gaps (*knowledge gap awareness*)
3. Focusing attention on the search for deeper patterns rather than surface characteristics of the problem (*deep feature recognition*)
4. Inducing germane processing of information (Leppink, Paas, Van Gog, van Der Vleuten, & Van Merrienboer, 2014) to resolve incongruity and uncertainty that is inherent in the problem-solving process (*cognitive load*)

There is an equally important affective and motivational aspect inherent in the design of PS-I. For instance, the initial problem-solving is expected to facilitate students' *curiosity* (Naylor, 1981) to learn targeted concepts after spending time grappling with a novel problem. The overall perceived *affect* (a collective term for describing feeling states like emotions and moods) too plays an important role in regulating cognition and behavior (Watson, Clark, & Tellegen, 1988). PS-I, in particular, because of integrating variant-invariant features and contrasting cases in the problem design to create an affective hook, can be expected to evoke *positive affect* in the form of surprise, interest and confusion. These knowledge emotions associated with thinking and comprehending (Silvia, 2009) motivate exploratory action that is needed to keep generating multiple solutions.

On the other hand, moderate levels of *negative affect* in the form of anger, disgust and contempt can also be posited to be prevalent in preparatory problem-solving because of its deliberately designed ill-structured nature (Sinha, 2021). As a hostility triad of emotions often experienced together (Izard, 1977), appropriately appraising the resulting negative evaluation of the task may encourage active behaviors to address problematic aspects and mollify the situation (Harmon-Jones, Price, Gable, & Peterson, 2014). Finally, with high expected failure rates, students can also be expected to experience discomfort as they endure an unpleasant task situation with no specific accuracy feedback (*cognitive dissonance*) – this level of induced dissonance (Festinger, 1962; Levin, Harriott, Paul, Zhang, & Adams, 2013) might in turn differentially affect problem-solving performance.

Prior PS-I research has empirically assessed the impact of these mechanisms across different study contexts such as mathematics (Likourezos & Kalyuga, 2017; Loibl & Rummel, 2014b; Newman & DeCaro, 2019), physics (Glogger-Frey, Gaus, & Renkl, 2017; Lamnina & Chase, 2019), medicine (Marei, Donkers, Al-Eraky, & Van Merrienboer, 2019) and data science (Sinha et al., 2020), primarily for high school students and undergraduates. However, their impact altogether thus far has not been examined within a single study context. We also do not know the extent to which these mechanisms are triggered in scaffolded PS-I contexts, in particular, in the presence of explicit scaffolding towards failure (cf. section 4.3.2). Additionally, novel mechanisms pertinent to students' metacognitive biases have not been explored within PS-I yet.

We posit that the highly generative nature of tasks within the initial problem-solving phase calls on students' metacognitive monitoring and regulation (Ackerman & Thompson, 2017) to think about progress. Students use this knowledge to adapt and make changes to their problem-solving strategies. These metacognitive judgments, often based on heuristic cues due to non-readily verifiable outcomes of solution revision, may not always reliably reflect actual knowledge. Such biases have been previously emphasized in the metacognitive literature (see Bjork, Dunlosky, and Kornell (2013) and Roebers (2017) for reviews). We are therefore also interested in investigating whether students' *calibration bias* (Kruger & Dunning, 1999), that is, the gap between performance evaluation and actual performance, could offer new insights into the impact of metacognitive awareness on learning within PS-I designs. Prior PS-I research (e.g., Loibl and Rummel (2014b) and follow-up work) has often used questionnaires tapping into the overarching construct of global awareness of knowledge gaps (that is, awareness without being able to identify which specific component is lacking) as an explanatory basis for the efficacy of PS-I designs. However, no prior work has directly measured metacognitive biases in close relation to problems solved during the preparatory phase and/or posttest.

The interplay between the aforementioned cognitive, affective, and meta-cognitive learning mechanisms can be posited to contribute to the pedagogical usefulness of PS-I, in particular in preparing students to learn, or facilitating their readiness to learn a targeted concept in follow up instruction (Schwartz & Martin, 2004). Capturing information about these mechanisms (via pre-planned measurements) is therefore critical to making claims about the explanatory basis of PS-I. Here, we use several such probes (cf. section 4.6.3) to explain why some scaffolded problem-solving experiences work better or worse than others.

2.3. Individual differences in learning from success and failure

Factoring individual differences in learning from failure and success (Clifford, 1984) is imperative in view of examining the spectrum of students who might differentially respond to scaffolding opportunities during problem-solving. Because a key preparatory goal in PS-I is to activate relevant prior knowledge (Kapur & Bielaczyc, 2012; Loibl et al., 2017), it is natural to account for domain-specific and task-specific *prior knowledge* that students use to generate and revise solutions. When a concept is not formally learned yet (e.g., in preparatory problem-solving), the associated cognitive demands are higher compared to problem-solving that is focused solely on the practice of already learned materials (Kapur, 2014; Likourezos & Kalyuga, 2017; Newman & DeCaro, 2019). This calls on students' *effort regulation* to steer through the task by exercising self-control and remaining focused (Pintrich, 1991).

Further, two motivational characteristics, *self-esteem* (Harter, 2012; von Soest, Wichstrom, & Kvalem, 2016) and *goal orientation* (Button, Mathieu, & Zajac, 1996), shape whether students view failures as opportunities to learn (Dweck, 1992), and more generally, affect their attributional style (Fielstein et al., 1985) towards failure and success (positive events to stable, global and internal causes, and negative events to temporary, specific or external causes). Finally, students' *attitude toward mistakes* (Leighton, Tang, & Guo, 2015) affects how they appraise the value of making mistakes, behaviors and affective reactions, all of which can enhance or impede receptivity to failures.

Better incoming characteristics, as evidenced by the aforementioned individual differences we measure in the current study (cf. section 4.6.1), can be assumed to positively influence responses to failure and success. For example, students with high self-esteem and a learning goal orientation disposition are likely to engage in deeper processing of information presented in the scaffold (Sinha et al., 2020). With sustained efforts toward meaning-making with the scaffold (high effort regulation), failure likelihood can be reduced.

3. The replication context

We aim to carry out a controlled experiment on the effects of explicit failure-driven and success-driven scaffolding on conceptual understanding and transfer. For the design of scaffolding, we draw from the seminal work on structuring and problematizing student production (Reiser, 2004). Structuring scaffolds increase success-likelihood by reducing degrees of freedom to lower task complexity, help students maintain direction, and make problem-solving tractable. Problematizing scaffolds increase failure likelihood by increasing degrees of freedom to challenge students' current understanding and highlight discrepancies between what students generate and the canonical task features. To contextualize our replication and extension, we first discuss the experiment reported by Sinha et al., 2020 and then summarize our list of major changes.

3.1. Classroom study by Sinha et al. (2020)

This study was conducted in data science education with $N = 221$ university students. The targeted learning concepts comprised introductory, but fundamental ideas in data science – Anscombe's Quartet (complementary importance of graphical + numerical representations in reasoning with data) and Spurious Correlation (correlation $\neq$ causation). The intervention comprised an initial problem-solving phase and a follow-up instruction phase, resembling PS-I (Loibl et al., 2017).

The type of scaffolding during the problem-solving phase was manipulated, resulting in four experimental conditions. Two variants of problematizing were used, each of which offered single-step scaffolds to nudge students towards reasoning with different suboptimal representations. These representations differed in their level of suboptimality with respect to the canonical answer. Further, two variants of structuring were used, each of which offered single-step scaffolds to nudge students towards reasoning with different optimal representations. These representations differed in their level of specificity with respect to closeness to the canonical answer. Posttest assessments comprised an isomorphic and a non-isomorphic conceptual understanding question.

Results demonstrated that failure-driven scaffolding was better than success-driven scaffolding with high specificity, however similar to success-driven scaffolding with low specificity. Further, students exposed to failure-driven scaffolding demonstrated a higher quality of constructive reasoning (meaningful elaborations (Chi, 2009) that went beyond what was presented), relative to success-driven conditions with both low and high specificity. These results were salient for the more complex Anscombe's Quartet topic.

3.2. Changes in the present study

We built on the quasi-experimental study by Sinha et al. (2020) to design a controlled experimental study. First, we improved the assessment of prior knowledge and posttest learning by administering a domain-general math ability calculus pretest (Epstein, 2007), and adding additional items assessing non-isomorphic conceptual understanding and transfer. Second, we designed multi-step scaffolds that progressively nudged students towards success or failure. In addition to design-level changes, we added a third experimental condition that provided no explicit scaffolding during the problem-solving phase (resembling a pure Productive Failure (Kapur & Bielaczyc, 2012) design). Finally, because previous results were salient for the Anscombe's Quartet topic, we focused on this topic.

3.3. Research questions and hypotheses

In light of prior work in PS-I and the discussed tradeoffs in scaffolding, the present study addresses the following research questions and their associated hypotheses.

1. *RQ1a:* How does scaffolding type during preparatory problem-solving (success-driven, failure-driven, none) impact students' *preparatory problem-solving performance*, controlling for their incoming characteristics and task experiences during the initial problem-solving phase?

Hypothesis 1a. Due to lower failure likelihood owing to a focus on revising and improving self-generated solutions and/or working with explicitly offered optimal representations, we expected students in the Success-driven and no scaffolding condition to have higher preparatory problem-solving performance than Failure-driven condition.

2. *RQ1b:* How does scaffolding type during preparatory problem-solving (success-driven, failure-driven, none) impact students' *post-test performance*, controlling for their incoming characteristics, task experiences during the initial problem-solving phase, and perceived lecture quality?

Hypothesis 1b. Due to the (a) differential overlap between the nature of preparatory student work with high success-likelihood the nature of work required to solve different posttests, and consequently (b) relatively lower opportunities for relevant learning mechanisms (cf. section 2.2) to be triggered in the absence of suboptimal representation generation, we expected Success-driven condition students to have

- the highest performance for *isomorphic* assessments (with answers corresponding precisely to the optimal representations offered during the problem-solving phase)
- similar performance as Failure-driven condition students for *non-isomorphic* assessments (with answers requiring a relatively higher depth of understanding and not depending exclusively on students' work with scaffolds, a trend also found in Sinha et al. (2020))
- the lowest performance for *transfer* assessments (with answers involving flexible integration of representations not covered during the intervention, and being more likely to be aided by failure-driven problem-space exploration)

3. *RQ2:* How do underlying mechanisms (*task experiences*) triggered because of engaging in preparatory problem-solving vary differentially for students receiving success-driven, failure-driven, or no scaffolding?

Hypothesis 2. Due to the relatively higher proportion of time on task where students work with self-generated and/or explicitly offered sub-optimal representations (and consequently, enhanced preparation for learning from instruction), we expected students in the Failure-driven and no scaffolding condition to have higher self-reported scores on measurements tapping facilitatory underlying mechanisms, relative to Success-driven condition students.

4. *RQ3:* What are the trends in *metacognitive calibration* across solutions developed during the problem-solving phase and posttest, and how do these vary for students receiving success-driven, failure-driven, or no scaffolding during preparatory problem-solving?

Hypothesis 3. Due to (a) greater opportunities for accurately self-evaluating problem-solving performance amidst exposure to self-generated and/or explicitly offered suboptimal representations, and consequently, owing to such practice, (b) an increased likelihood of awareness of representations that may not provide a clear canonical solution pathway, we expected students in the Failure-driven condition and no-scaffolding condition to be relatively better calibrated than Success-driven condition students.

4. Method

4.1. Participants

We recruited $N = 132$ participants (59% male, $n = 78$; 41% female, $n = 54$) from a large student volunteer pool (>10,000) from two highly-ranked but open-admission universities in Europe. Apriori power analysis based on ANCOVA suggested sample size to be in the range of $N = 128–206$ to detect medium-size effects ($d = 0.5$) with 70%–90% power (α error probability 0.05). To participate in the study, students had to know high school math (basic algebra, calculus, statistics and probability) and be familiar with programming in Python (at least 1 semester of Python programming experience).[1] 58.33% of participants in our sample came from ten different European ethnicities (30.3% German, being the majority), and 41.67% of participants came from a non-European ethnicity (37.12% Asian, being the majority).

Post-study questionnaires revealed that the majority (83.3%) of participants had not heard of Anscombe's Quartet prior to the study session, suggesting that they were novices in the targeted concept (no task-specific knowledge). However, participants reported that they were somewhat familiar with numerical representations ($M = 4.15$, $SD = 1.77$) and graphical representations ($M = 4.48$, $SD = 1.68$) before coming into the study (on a 7-point Likert scale). This suggests that students might have possessed some prerequisite domain-specific knowledge for learning the targeted concept.

4.2. Task domain

Anscombe's Quartet, where the aim was that students understand the complementary importance of numerical and graphical representations when reasoning with data, was the targeted data science learning concept. To work towards this goal, a set of datasets with similar descriptive statistics but very different plots were presented to students, giving them the opportunity to reflect on a concrete task that relied on assessing the strength of evidence from both these data sources. Students had to work individually in a dynamically executable online problem-solving environment (Jupyter notebook), similar to the one used in the classroom study by Sinha et al. (2020). Using form fields and interactive sliders, this dynamically executable Python Jupyter notebook allowed students to load relevant datasets for a problem, run basic statistics, read information about the problem-solving task, record their answers, reasoning and confidence (see Fig. 1). Similar to a programming language compiler, the Python Jupyter notebook provided syntax-level feedback when writing programming code. Posttest questions were also administered using separate Python Jupyter notebooks.

[1] The chosen high school math prerequisites were kept in mind when designing the learning materials for the current study as well as in our previous study (Sinha et al., 2020). Because of iterative piloting of these learning materials with university students prior to the actual study, we assume that the criterion of knowing high school math should be fulfilled by all participants. To maximize participation, we intentionally kept this statement in the recruitment advert to reflect that the bar for signing up was low and that no special/-advanced math skills were needed. Further, students were invited for participation if they scored >= 7/10 on a pre-screening quiz covering Python syntax.

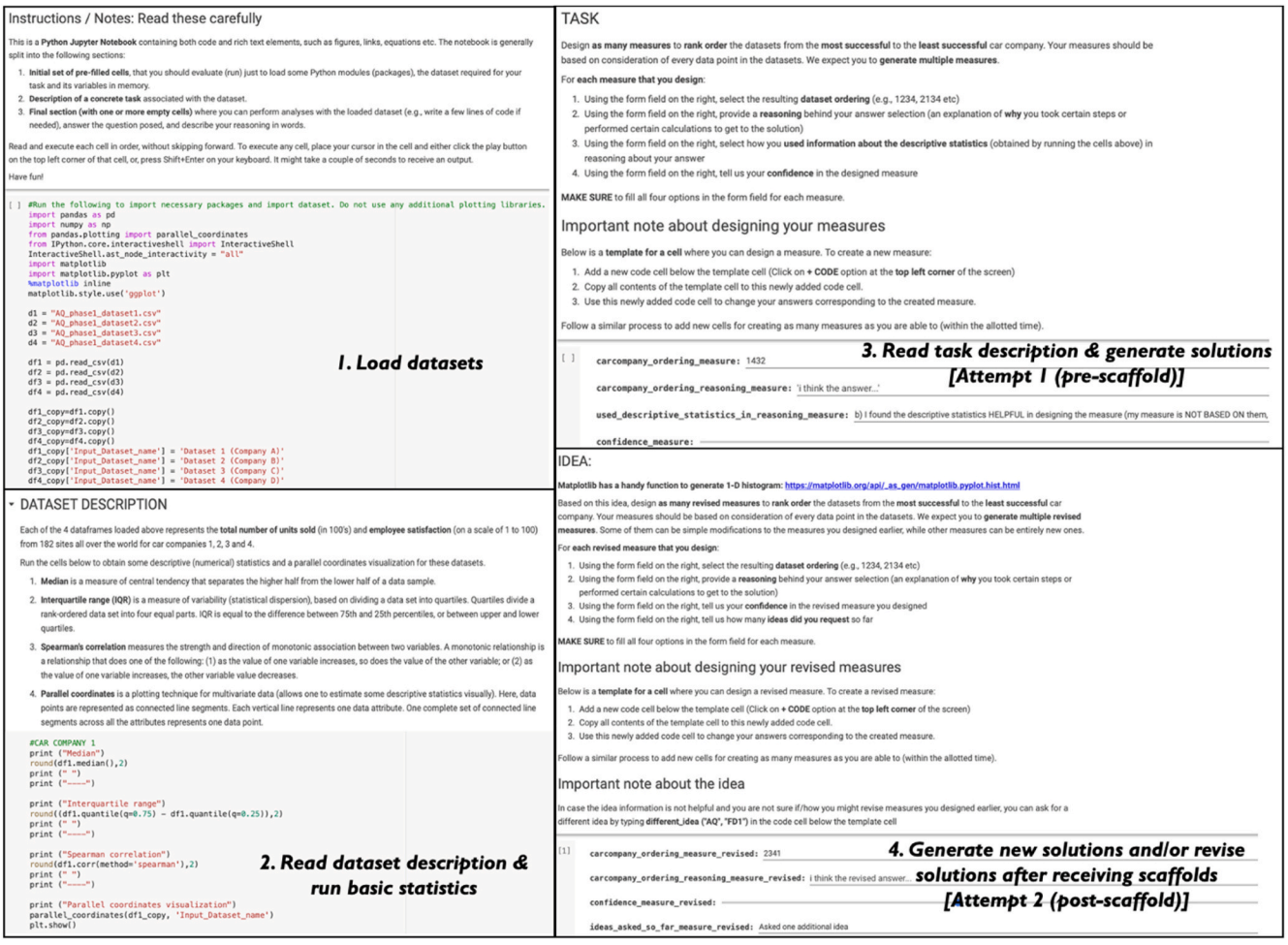


Fig. 1. Problem-solving environment used in the study.

4.3. Scaffolding design

4.3.1. Success-driven scaffolding

Our preparatory problem-solving task involved reasoning with a bivariate dataset. Therefore, the structuring scaffold hierarchy included a prompt (a Wikipedia page suggesting students to read more about exploratory data analysis), hint (description of data science phenomena under consideration), and finally a bottom-out hint or the last hint in the sequence precisely conveying the answer (syntax for scatterplot generation, an optimal graphical representation for reasoning with bivariate datasets). *Prompts* divulged very little solution-relevant information and can be conceived as the weakest or least-specific structuring scaffold (smallest nudge towards success). Prompts pointed students to problem conditions that should likely remind students of the knowledge components' relevance. *Hints*, on the other hand, incorporated the idea of teaching students the knowledge component that is actually relevant in the current problem-solving context. This means that hints told students what to do but not how. Finally, *bottom-out hints* represented the strongest or most-specific structuring scaffold (biggest nudge towards success) and told precise and potentially optimal ways of proceeding.

Presentation-wise, students first received a prompt in their problem-solving workbooks. Further scaffolds such as the hint and bottom-out hint were only revealed if students indicated that the information from a previous scaffold was not helpful in solution revision, and explicitly asked for the next scaffold. The underlying design rationale was to give students the least-specific structuring that could nudge problem-solving towards success first. This was based on empirical effectiveness of such a scaffold hierarchy (Aleven, McLaughlin, Glenn, & Koedinger, 2016) based on the behavior of expert human tutors (VanLehn, 2011), and used in prominent educational applications (e.g., Khan Academy, Carnegie Learning, ASSISTments).[2]

4.3.2. Failure-driven scaffolding

Our problematizing scaffold hierarchy started with the presentation of a moderately high suboptimal representation (one-dimensional histogram), subsequently an extremely high suboptimal representation (bar chart), and finally ended with the least suboptimal representation among the three (two-dimensional histogram). A bar chart is suboptimal because it is only informative when variables in the data comprise a mix of continuous and categorical variables. Moreover, even if one makes a bar chart with only numeric variables, there is high noise in a bar chart visualization in absence of natural ordering of the dataset categories – consequently, it is difficult to discover clear patterns when comparing arbitrary data segments. With histograms, information is lost because of binning and/or the lack of directly perceivable information about covariation in the data. Such effects are more pronounced for one-dimensional histograms compared with two-dimensional histograms.

As with success-driven scaffolds, students could request these scaffolds one at a time, if they were not certain about revising their answers. Our design rationale behind the failure-driven scaffold hierarchy, where initially presented scaffolds were more suboptimal than the ones

[2] *Khan* Academy (https://www.khanacademy.org/), Carnegie Learning Inc. (https://www.carnegielearning.com/products/software-platform/mathia-learning-software/), ASSISTments (https://new.assistments.org/).

presented later, was to increase the likelihood that students use all scaffolds and explore the problem-space maximally, instead of following an isolated solution path.[3] We intended to lead students towards questionable decision-making by asking them to consider a subset of conceptual domain factors that did not lead to the canonical solution and challenge them to reason with such partially-gained insights. To keep scaffold presentation style consistent with the Success-driven condition and avoid any bias before students interacted with the scaffolds, we kept a relatively neutral tone and introduced each failure-driven scaffold as an idea.

4.3.3. No explicit scaffolding

We also compared explicit success-driven and failure-driven scaffolding to a PS-I learning design without any explicit scaffolding during the problem-solving phase. This resembles a pure Productive Failure condition (Kapur, 2014; Kapur & Bielaczyc, 2012), and affords a direct evaluation of the efficacy of initial scaffolding and its preparatory benefits in learning from instruction.

4.4. Learning materials

All learning materials underwent iterative design and testing with participants having similar demographic as the study ($N = 20$), and can be found under supplementary materials.

4.4.1. Problem-solving phase

In the problem-solving phase, students had to create as many measures as they could, in order to rank-order datasets of four car companies from the most successful to least successful. Information about the number of car units sold and employee satisfaction was given to them. Consistent with the design principles of Productive Failure (Kapur & Bielaczyc, 2012), we designed the datasets such that they had exact same non-parametric statistics (median, interquartile range, Spearman's correlation), but different parametric statistics (mean, standard deviation, Pearson's correlation) and very different visualizations. For the second attempt workbook, students in the Success-driven and Failure-driven conditions received multi-step structuring or problematizing scaffolds. They could use information from these scaffolds to create as many revised solutions as they could. In the no scaffold (Productive Failure) condition, students did not receive any explicit scaffolds in the second attempt. They were asked to keep generating more solutions.

4.4.2. Instruction phase

The instruction phase was presented as a three-part video that students could watch at their own pace (e.g., by pausing, re-watching). The first part introduced the concept of Anscombe's Quartet (5 min), the second part compared and contrasted different student solutions with the canonical one (10 min), and the third part presented and justified one possible canonical answer to the task (5 min). The lecture content was developed by the first author with more than 5 years of data science experience, and further independently verified by two instructors with backgrounds in data science and statistics.

4.4.3. Posttest

Our posttest included one isomorphic item testing for conceptual understanding, two non-isomorphic items testing for conceptual understanding, and one item testing for transfer (item 4). Final scoring was binary (0/1) for all items. The *isomorphic* conceptual understanding question (item 1) involved reasoning with both graphical and numerical representations in a task that had a two-dimensional data distribution (similar to the problem-solving phase, however, with an entirely different cover story). The first *non-isomorphic* conceptual understanding question (item 2) involved reasoning with one-dimensional data distribution, as opposed to a two-dimensional data distribution used in the initial problem-solving task. The second *non-isomorphic* conceptual understanding question (item 3) involved reasoning with a tri-variate dataset and making inferences based on the descriptive technique of linear regression. The *transfer* question (item 4) involved reasoning with not only descriptive statistics but also inferential statistics that were not covered during the problem-solving task. Details can be found in supplementary materials.

4.5. Experimental design

The complete intervention took 150 min (see Fig. 2). Prior to the problem-solving phase, students' incoming characteristics were assessed using a combination of questionnaires and a short calculus pretest (20 min, cf. section 4.6.1). Participants were randomly assigned to experimental conditions. During the initial problem-solving phase, students in every condition made an identical first attempt at the task in the absence of any external scaffolds (20 min). For problem-solving in the second attempt, three experimental manipulations were instantiated (20 min). We call these Failure-driven ($N = 45$, explicit problematizing scaffolds nudging students towards failure), Productive Failure ($N = 43$, no explicit scaffolds), and Success-driven ($N = 44$, *explicit* structuring scaffolds nudging students towards success). After the problem-solving phase, we collected students' task experiences using questionnaires (5 min, cf. section 4.6.3).

After reporting their problem-solving task experiences, students in all conditions went through an identical instruction phase (20 min), delivered in the form of a pre-recorded video lecture using the Go-Lab infrastructure (De Jong, Sotiriou, & Gillet, 2014). Finally, students solved four posttest questions tapping different dimensions of understanding of the targeted concept (60 min). After completion of the posttest, students' perceived lecture quality was captured using questionnaires. In addition, students also self-reported their demographic information (gender, ethnicity), high school math score, and familiarity with numerical/graphical representations (5 min, cf. section 4.6.6).

4.6. Measures

4.6.1. Before the problem-solving phase

Students' *incoming characteristics* were collected via questionnaires (randomly ordered during presentation), which previous work (Sinha et al., 2020) has validated. These characteristics reflect individual differences in learning from success and failure (Clifford, 1984). We included measurements of effort regulation (4 items, $\alpha = 0.74$, e.g., *"I work hard to do well even if I don't like what we are doing in classes"*), goal orientation (learning (8 items, $\alpha = 0.84$, e.g., *"The opportunity to do challenging work is important to me"*) and performance (8 items, $\alpha = 0.71$, e.g., *"I prefer to do things that I can do well rather than things that I do poorly"*) sub-scales), attitude towards mistakes (affect (4 items, $\alpha = 0.67$, e.g., *"When I make mistakes answering classroom questions, I am overwhelmed with*

[3] We chose not to have a failure-scaffold hierarchy based on the principle of *smallest to biggest nudge towards failure* (two-dimensional to one-dimensional histogram to bar chart), because the presentation of a scaffold with relatively low suboptimality at the beginning may already start pushing students towards a reasonable answer (inferences about the relationship between two numeric variables). This, in turn, may reduce the likelihood that students move onto the exploration of more suboptimal scaffolds. We also chose not to present failure-scaffolds following the principle *biggest to smallest nudge towards failure* (bar chart to one-dimensional histogram to two-dimensional histogram), because it can in fact be perceived to resemble a structuring/success-driven sequence that makes critical task features for solving the problem increasingly more relevant. Additionally, based on the help-seeking literature in intelligent tutoring systems (Aleven, Stahl, Schworm, Fischer, & Wallace, 2003), this sequencing posed a risk of help-abuse, that is, students just clicking through to the last scaffold that they would probably perceive to be the most useful (here, a two-dimensional histogram, which is comparatively the least suboptimal representation) – in such a case, the likelihood of meaningful engagement with all scaffolds would be reduced. The failure-driven scaffold hierarchy implemented in our study served as a middle ground between these two options.

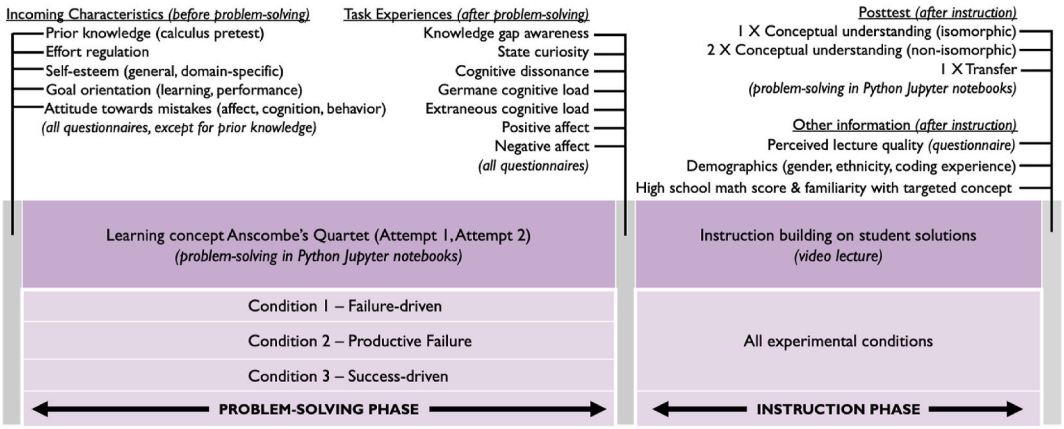


Fig. 2. Experimental design of the study. For students who were selected to participate in the study, depicted measures were administered at three time points (before problem-solving phase, after problem-solving phase, after instruction) during the study.

embarrassment"), cognition (4 items, α = 0.66, e.g., *"I believe it is smart to avoid making mistakes during learning"*) and behavior (6 items, α = 0.73, e.g., *"When I make mistakes on an exam, I feel motivated to study harder"*) sub-scales), self-esteem (general (5 items, α = 0.77, e.g., *"Some university students are often disappointed with themselves BUT Other university students are pretty pleased with themselves (reversed)"*) and domain-specific (5 items, α = 0.66, e.g., *"Some university students do very well at their classwork BUT Other university students don't do very well at their classwork"*) sub-scales), and prior knowledge (calculus pretest[4] (8 questions, α = 0.71)). Prior knowledge assessment also included post-study measurements of high school math scores and familiarity with numerical and graphical representations. Table 1 provides operational definitions. Complete scales are present in supplementary materials.

4.6.2. Problem-solving phase

For each problem-solving attempt, *solution quantity* was computed by simply counting the number of unique solutions students developed in the problem-solving phase. The computation for *solution quality* accounted for how many rank-ordered pairs were correctly identified (scores ranged from 1 to 6). We focused on the best solution that each student generated, that is, the solution with the highest score.[5] For each attempt, students also reported their confidence for the solutions developed using a 5-point slider. Empirical assessments of students' monitoring processes were captured by looking at the gap between solution quality and reported confidence.

4.6.3. After the problem-solving phase

Students' *task experiences* were collected via questionnaires (randomly ordered during presentation). These experiences tap onto different cognitive and affective mechanisms (Kapur, 2016; Loibl et al., 2017) posited to attribute to the preparatory benefits of PS-I. Previous empirical work in PS-I has validated self-reported measurements for these task experiences, including knowledge gap awareness (Sinha et al., 2020; Glogger-Frey et al., 2017; Loibl & Rummel, 2014b; Newman & DeCaro, 2019), state curiosity (Sinha et al., 2020; Lamnina & Chase, 2019; Loibl & Rummel, 2014b), cognitive load (Sinha et al., 2020; Glogger-Frey et al., 2017; Likourezos & Kalyuga, 2017; Marei et al., 2019; Newman & DeCaro, 2019), affect (Lamnina & Chase, 2019) and cognitive dissonance (Sinha et al., 2020), which lends credibility to their use in the current study.

Accordingly, we draw on this established PS-I work and include measurements of knowledge gap awareness (5 items, α = 0.81, e.g., *"My knowledge was insufficient to carry out these tasks"*), cognitive dissonance (6 items, α = 0.53, e.g., *"Some of the answers I gave in these tasks were inconsistent with my previous beliefs about the topics"*), state curiosity (9 items, α = 0.86, e.g., *"I feel like asking questions about what is happening"*), germane cognitive load (6 items, α = 0.83, e.g., *"This activity improved my understanding of the content that was covered"*), extraneous cognitive load (4 items, α = 0.7, e.g., *"The explanations, instructions and clues in this activity were full of unclear language"*), positive affect (10 items, α = 0.91, e.g., determined, enthusiastic), and negative affect (10 items, α = 0.89, e.g., upset, distressed). Table 1 provides operational definitions. Complete scales are present in supplementary materials.

4.6.4. Instruction phase

We recorded the time that students spent on watching each of the three parts of the video lecture (in seconds).

4.6.5. Posttest

Varimax-rotated principal component analysis was used to reduce the correlated (binary) posttest *scores* to a smaller set of important independent composite scores. Further, a coding scheme, based on prior work (Chi, 2009; Kapur & Kinzer, 2009) and validated in Sinha et al. (2020), was applied to quantify *reasoning quality*, or more specifically, the average percentage of complete mathematical and non-mathematical elaborations comprising graphical and/or numerical representations. Coding was conducted blind to the experimental condition. Details of the coding scheme for computing this percentage score can be found in supplementary materials. Students also reported confidence in their answers using a 5-point slider. Empirical assessments of students' monitoring processes were captured by looking at the gap between posttest scores and reported confidence.

4.6.6. After the posttest

We used an instructional skills questionnaire (randomly ordered during presentation) to capture students' perceived *lecture quality* along six dimensions. The questionnaire, adapted from Knol, Dolan, Mellenbergh, and van der Maas (2016), assessed structure (4 items, α = 0.81, e.g., *"The instructor gives clear summaries"*), explication (4 items, α = 0.82, e.g., *"The instructor's explanations are hard to follow (reversed)"*), stimulation (4 items, α = 0.77, e.g., *"The instructor enlivens the subject matter"*), validation (4 items, α = 0.73, e.g., *"The utility of the subject matter is hardly discussed (reversed)"*), instruction (4 items, α = 0.74, e.g., *"The instructor indicates*

[4] The original German version of the calculus pretest (Epstein, 2007) comprised 11 items and was translated to English for our study. Based on initial piloting with $N = 20$ students, we dropped two items that were negatively correlated with the total scale. An additional item was removed from all analyses due to a translation error in wording.

[5] As students generate more solutions, their average solution quality is expected to decrease. Taking solution quality to be representative of the best solution prevents this confound of solution quality and quantity.

Table 1
Operational definitions of measures used in the study.

Measure	Definition
Incoming characteristics (administered before the problem-solving phase)	
Effort regulation	the ability to control effort and attention when tasks are difficult (Pintrich, 1991)
Goal orientation	a dispositional trait, which makes one more likely to focus (or lack, thereof) on viewing failures as opportunities to learn (Button et al., 1996)
Learning goal orientation	desire to perform challenging work, learn new skills, develop alternative strategies when working on difficult task
Performance goal orientation	desire to avoid negative judgments of one's competence
Attitude towards mistakes	a relatively enduring organization of beliefs, feelings, and behavioral tendencies towards mistakes (Leighton et al., 2015)
Affect	affective reactions (emotional, physiological experiences) when making mistakes
Cognition	the perceived utility of making mistakes (beliefs reflecting a positive or negative evaluation of mistakes)
Behavior	observable actions undertaken to avoid or embrace mistakes
Self-esteem	perceived competence, which positively affects attributional style towards success and failure (Harter, 2012)
General self-esteem	a general perception of the self
Domain-specific self-esteem	perceived scholastic cognitive competence, as applied to university work
Task experiences (administered after the problem-solving phase)	
Knowledge gap awareness	the extent to which students realize what they do not know (Glogger-Frey et al., 2017)
Cognitive dissonance	a state of discomfort associated with detection of conflicting concepts (Levin et al., 2013)
State curiosity	desire to know more to fill the perceived knowledge gaps (Naylor, 1981)
Germane cognitive load	load from relating relevant information from long-term memory to new information elements (Leppink et al., 2014)
Extraneous cognitive load	load from engaging in processes that may not contribute directly to construction of cognitive schemata (Leppink et al., 2014)
Positive affect	the extent to which an individual subjectively experiences positive moods such as joy, interest, and alertness (Watson et al., 1988)
Negative affect	feelings of emotional distress, defined by the common variance between anxiety, sadness, fear, anger, and other unpleasant emotions (Watson et al., 1988)
Perceived lecture quality (administered after the posttest)	
Structure	extent to which the instructor handled the subject matter systematically and in an orderly way
Explication	extent to which the instructor explained the subject matter, especially the more complex topics
Simulation	extent to which the instructor raised interest in the subject matter
Validation	extent to which the instructor stressed benefits and the relevance of the subject matter for educational goals
Instruction	extent to which the instructor provided instructions about how to study the subject matter
Activation	extent to which the instructor encouraged actively thinking about the subject matter

which parts of the subject matter are essential"), and activation (2 items, *Spearman-Brown* coefficient = 0.64, e.g., *"Students are encouraged to think along during the lecture"*).[6] Table 1 provides operational definitions. Complete scales are present in supplementary materials.

4.7. Analysis plan

Before carrying out the analysis for our stated research questions, we performed a manipulation check for the fidelity of scaffolding implemented during problem-solving, and for the fidelity of the delivered lecture.

4.7.1. RQ1 and RQ2

To account for the multivariate nature of our dependent variables for research questions 1 and 2, MANCOVAs were run first to examine omnibus effects across the Failure-driven, Productive Failure, and Success-driven experimental conditions. Subsequently, we ran univariate ANCOVAs, along with Bonferroni-corrected posthoc t-tests (pairwise comparisons adjusted for a family of 3) between the experimental conditions. Running these follow-up pairwise t-tests also allowed us to obtain the effect size (Cohen's *d*) estimates for each dependent variable.

To assess how the conditions differentially impacted performance during the preparatory problem-solving phase (RQ1a), solution quantity (attempt 1 and attempt 2) and solution quality (attempt 1 and attempt 2) were used as dependent variables. We controlled for students' incoming characteristics[7] and task experiences during the problem-solving phase. An analogous analysis of the posttest (RQ1b) included the score (isomorphic, non-isomorphic, transfer) and reasoning quality (isomorphic, non-isomorphic, transfer) as dependent variables. We controlled for students' incoming characteristics, task experiences during the problem-solving phase, and perceived lecture quality. Finally, the analysis for RQ2 focused on differences in the underlying mechanisms (cf. section 4.6.3) across conditions and controlled for students' incoming characteristics. Because the dependent variables for RQ1a, RQ1b and RQ2 were measured at different time points during the design, not all sets of covariates were relevant for each RQ – they are represented accordingly with tick marks in Table 2.

The relatively small sample sizes also led us to focus attention on Cohen's *d* measure of effect size (along with 95% confidence intervals) to evaluate the effectiveness of our scaffolded PS-I intervention. For judging the practical importance of standardized effect size estimates, Hill, Bloom, Black, and Lipsey (2008) recommend the use of empirical benchmarks of comparison that reflect the nature of the intervention being evaluated, its target population, and the outcome measure or measures being used. Our recent meta-analytic work comparing PS-I with scaffolded PS-I designs (Sinha & Kapur, 2021b) has found the average effect size disfavoring PS-I to be Hedge's g −0.08 [95% CI -0.34, 0.28], for a similar student population (undergraduate level) and

[6] The original version of the activation subscale for the perceived lecture quality questionnaire had 4 items (Knol et al., 2016). In the current study, however, we did not use 2 of the items that focused on providing discussion opportunities during the lecture ("*The instructor provides little opportunity for discussions*" and "*During this lecture there is hardly any occasion to discuss the subject matter*"). This was due to the video-based (rather than in-person) nature of the lecture. We therefore also used the Spearman-Brown prophecy formula (Spearman, 1910) to calculate the predicted reliability of the activation subscale, given the original reliability and an expansion of the scale to 4 items. The predicted reliability value was 0.78.

[7] Despite 16.7% of participants having heard about the targeted learning concept prior to the intervention, we did not include this binary variable as a covariate in our primary analyses for RQ1 and RQ2. This was because of concerns regarding the validity of this variable as a representative for prior knowledge, exemplified in its low correlations with self-reported familiarity with numerical representations (r_{pb} = 0.26, p = 0.003) and graphical representations (r_{pb} = 0.19, p = 0.031). Inclusion of this additional covariate, however, did not change the trends in reported results.

Table 2
Summary of the analysis plan for the study.

	Dependent Variables	Covariates		
		Incoming characteristics	Task experiences	Lecture quality
RQ1a	Solution quantity (Attempt 1, Attempt 2)	✓	✓	
	Solution quality (Attempt 1, Attempt 2)	✓	✓	
RQ1b	Score (Isomorphic, Non-isomorphic, Transfer)	✓	✓	✓
	Reasoning (Isomorphic, Non-isomorphic, Transfer)	✓	✓	✓
RQ2	Knowledge gap awareness	✓		
	Cognitive dissonance	✓		
	State Curiosity	✓		
	Germane/Extraneous cognitive load	✓		
	Positive/Negative affect	✓		
RQ3	Calibration: Δ(Confidence, Attempt 1 Solution Quality)		not applicable	
	Calibration: Δ(Confidence, Attempt 2 Solution Quality)		not applicable	
	Calibration: Δ(Confidence, Isomorphic Score)		not applicable	
	Calibration: Δ(Confidence, Non-isomorphic Score)		not applicable	
	Calibration: Δ(Confidence, Transfer Score)		not applicable	

Note. For RQs 1 and 2, MANCOVA (univariate ANCOVAs) and post-hoc Bonferroni-corrected pairwise comparisons (t-tests) were used. For RQ 3, a one-sample *t*-test was used to assess differences from 0.

learning outcomes (conceptual understanding, transfer) such as ours. These meta-analytic effect size results led us to consider all effect sizes $> |0.2|$ as a benchmark for improvement of scaffolded PS-I (that is, the Success-driven and Failure-driven conditions) over unscaffolded PS-I (that is, the Productive Failure condition) in the current study.

As alternative complementary analyses for null Hypothesis significance testing (NHST),[8] we also used a Bayesian informative hypotheses evaluation ANCOVA (Hoijtink, Mulder, van Lissa, & Gu, 2019) for RQ1 and RQ2. This allowed direct evaluation of the Bayes factor of a specific hypothesis at hand versus its complement (BF_c).[9] For any particular dependent variable, these hypotheses were based on the actual descriptive trends in marginal means that we observed across the experimental conditions – marginal means, which we obtained after running the regular frequentist ANCOVA. As an example, the complement of a hypothesis *H* ($\mu1 > \mu2 > \mu3$ for the Failure-driven, Productive Failure and Success-driven conditions) would comprise any set of restrictions between the means that is not *H*.

4.7.2. RQ3

To assess calibration (RQ3), we used a one-sample *t*-test to compare the difference (Δ) between scaled values of confidence judgments and performance (that is, solution quality of the best solution developed for the problem-solving phase, and scores for the posttest). We assessed whether this difference was significantly different from 0, indicating under-confidence ($\Delta < 0$) or over-confidence ($\Delta > 0$) bias. A non-significant difference (with low effect sizes) would indicate students to be metacognitively well-calibrated.

[8] We used non-Bayesian analyses (NHST) because it is a dominant tool in psychological research, easily understood by the majority of the readers, and allows the computation of effect size (Cohen's *d*) estimates that can be compared with previous studies. However, given the criticisms regarding NHST in the past decade (Cumming, 2014), we further used a Bayesian alternative to NHST that allowed us to quantify the strength of evidence (uncertainty) in favor of a particular Hypothesis using the observed data (as opposed to a dichotomous reject/fail-to-reject decision that comes with null hypothesis significance testing, where Type I and Type II errors are determined independently of the observed data). We conceive of Bayesian informative hypotheses evaluation ANCOVA (Hoijtink et al., 2019) as a valuable alternative in the context of comparing experimental designs, and believe that with greater reporting of (and consequently, familiarity with) these more advanced Bayesian methods, their exclusive reporting might be more warranted.

[9] BF_c can be interpreted as odds in favor of an informative Hypothesis being tested relative to its complement. BF_c values can be interpreted as follows: 1–3 (weak/anecdotal), 3–10 (positive/substantial), 10–20 (positive/strong), 20–30 (strong), 30–100 (strong/very strong), 100–150 (strong/decisive), > 150 (very strong/decisive).

5. Results

Results from our implementation fidelity check during the problem-solving phase provided empirical evidence that Productive Failure and the two scaffolding conditions worked as intended, both in terms of (a) how they affected solution generation in attempt 2 (quantity and quality), and (b) how responses to presented scaffolds were related to solution quality improvements in attempt 2. Further, results from the implementation fidelity check also provide empirical evidence that the instruction phase, which was kept the same for all conditions as part of the experimental design, was judged to be equally good (average scores across most dimensions of perceived lecture quality were > 4 on a 7-point Likert scale). We provide an extensive elaboration of implementation fidelity in the supplementary materials.

5.1. Performance (RQ1)

5.1.1. Problem-solving phase (RQ1a)

The MANCOVA for solution quantity was significant, Wilks' $\lambda = 0.8$, $F(4, 162) = 4.86$, $p = 0.001$. Univariate ANCOVAs showed significant differences in solution quantity for attempt 2, $F(2, 105) = 5.06$, $p = 0.008$, $\eta_p^2 = 0.11$. *Solution quantity* in attempt 2 was higher for students in the Productive Failure condition compared to the Success-driven ($d = 0.61$) and Failure-driven ($d = 0.90$) conditions. Further, we also found students in the Success-driven condition to have a higher solution quantity ($d = 0.30$), relative to the Failure-driven condition. BF_c corresponding to this overall observed trend in attempt 2 solution quantity was 18.47.

Neither the MANCOVA nor univariate ANCOVAs (and Bonferroni-corrected follow-up t-tests) were significant for solution quality. However, the *solution quality* was descriptively higher in the Productive Failure ($d = 0.54$) and Success-driven condition ($d = 0.35$) for the first problem-solving attempt, relative to the Failure-driven condition. This was despite the initial randomization to experimental conditions and identical task instructions. BF_c corresponding to this overall observed trend in attempt 1 solution quality was 8.87. Similar trends existed in solution quality during attempt 2 with weaker odds, but still lying within the same positive/substantial Bayes Factor range of 3–10 ($BF_c = 3.86$)[10].

[10] After additionally controlling for differences in the first problem-solving attempt, the overall trends in the pattern of results (marginal means and standard errors) for solution quality in the second problem-solving attempt did not change. Similarly, posthoc pairwise t-tests between the experimental conditions revealed no major differences in the overall trends of effect size (Cohen's *d*) estimates. Taken together, this suggests that even after controlling for differences in the initial generation, the impact of scaffolding on immediate problem-solving performance still aligned with Hypothesis 1a.

Table 3
Problem-solving phase performance (RQ1a), posttest performance (RQ1b) and underlying mechanisms (RQ2) across experimental conditions.

	Marginal Mean (SE)			BF_c
	Failure-driven	Productive Failure	Success-driven	
Problem-solving phase [Attempt 1] [Pre-scaffold]				
Quantity	2.02 (0.23)	2.12 (0.25)	2.28 (0.24)	2.77
Quality (Max 6)	3.85 (0.18)	4.43 (0.2)	4.27 (0.19)	8.87
Problem-solving phase [Attempt 2] [Post-scaffold]				
Quantity	1.75 (0.27)	3.16 (0.31)	2.11 (0.28)	18.47
Quality (Max 6)	3.96 (0.23)	4.36 (0.26)	4.19 (0.24)	3.86
Posttest (Isomorphic conceptual understanding)				
Score (Max 1)	0.60 (0.08)	0.73 (0.08)	0.78 (0.07)	6.08
Reasoning (Max 100)	44.79 (5.51)	40.61 (5.95)	31.57 (5.37)	7.12
Posttest (Non-isomorphic conceptual understanding)				
Score (Max 1)	0.49 (0.08)	0.43 (0.08)	0.55 (0.07)	3.74
Reasoning (Max 100)	20.97 (3.86)	15.56 (4.09)	17.5 (4.09)	3.09
Posttest (Transfer)				
Score (Max 1)	0.60 (0.08)	0.39 (0.07)	0.47 (0.07)	8.38
Reasoning (Max 100)	22.91 (5.53)	17.54 (5.78)	19.75 (4.99)	2.32
Problem-solving Task Experiences				
Knowledge gap awareness (Max 5)	3.98 (0.13)	3.5 (0.14)	3.96 (0.13)	5.82
Cognitive dissonance (Max 5)	3.13 (0.09)	3.19 (0.09)	2.98 (0.09)	5.90
State curiosity (Max 5)	3.82 (0.12)	3.49 (0.13)	3.61 (0.12)	8.38
Germane cognitive load (Max 5)	2.60 (0.13)	2.52 (0.14)	2.35 (0.13)	5.03
Extraneous cognitive load (Max 5)	3.54 (0.12)	3.39 (0.13)	3.61 (0.13)	4.37
Positive affect (Max 5)	2.57 (0.12)	2.52 (0.13)	2.30 (0.13)	5.12
Negative affect (Max 5)	2.30 (0.13)	1.88 (0.14)	2.12 (0.13)	13.79

Note. For each row, marginal means and standard errors are depicted. Also depicted is the BF_c or the Bayes factor of the informative hypotheses (based on the marginal mean trends) versus its complement.

Taken together, the aforementioned results for solution quantity and solution quality support Hypothesis 1a regarding students in the Success-driven and Productive Failure conditions having higher preparatory problem-solving performance than students in the Failure-driven condition. Tables 3 and 4 summarize these results.

5.1.2. Posttest (RQ1b)

The rotated factor loadings of a 3-component principal component analysis (eigenvalues 1.44, 1.04, and 0.85), which accounted for 83% of the total variance, were in line with our intended differentiation between *non-isomorphic* conceptual understanding (items 2 and 3), *transfer*

Table 4
Effect size estimates for post-hoc pairwise comparisons (RQ1a, RQ1b, RQ2) between experimental conditions.

	Cohen's *d* [95% CI]		
	Failure-driven & Productive Failure	Failure-driven & Success-driven	Productive Failure & Success-driven
Problem-solving phase [Attempt 1] [Pre-scaffold]			
Quantity	−0.08 [-0.50, 0.34]	−0.18 [-0.62, 0.25]	−0.10 [-0.54, 0.33]
Quality	**−0.54 [-0.97, -0.11]**	**−0.35 [-0.78, 0.08]**	0.14 [-0.29, 0.57]
Problem-solving phase [Attempt 2] [Post-scaffold]			
Quantity	**−0.90 [-1.38, -0.41]**	**−0.30 [-0.76, 0.16]**	**0.61 [0.11, 1.10]**
Quality	**−0.31 [-0.77, 0.16]**	−0.17 [-0.62, 0.29]	0.12 [-0.36, 0.60]
Posttest (Isomorphic conceptual understanding)			
Score	**−0.31 [-0.83, 0.21]**	**−0.50 [-1.07, 0.09]**	−0.14 [-0.62, 0.34]
Reasoning	0.12 [-0.33, 0.58]	**0.40 [-0.06, 0.86]**	**0.30 [-0.15, 0.75]**
Posttest (Non-isomorphic conceptual understanding)			
Score	0.16 [-0.35, 0.68]	−0.16 [-0.66, 0.34]	**−0.34 [-0.82, 0.15]**
Reasoning	**0.24 [-0.21, 0.70]**	0.15 [-0.31, 0.62]	−0.10 [-0.57, 0.37]
Posttest (Transfer)			
Score	**0.56 [0.03, 1.08]**	**0.35 [-0.15, 0.85]**	**−0.21 [-0.69, 0.27]**
Reasoning	**0.21 [-0.31, 0.74]**	0.11 [-0.40, 0.63]	−0.09 [-0.60, 0.43]
Problem-solving Task Experiences			
Knowledge gap awareness	**0.53 [0.10, 0.95]**	0.03 [-0.38, 0.45]	**−0.52 [-0.94, -0.09]**
Cognitive dissonance	−0.09 [-0.50, 0.33]	**0.27 [-0.14, 0.69]**	**0.36 [-0.06, 0.79]**
State curiosity	**0.41 [-0.01, 0.83]**	**0.27 [-0.15, 0.69]**	−0.14 [-0.56, 0.28]
Germane cognitive load	0.09 [-0.33, 0.51]	**0.31 [-0.11, 0.73]**	0.20 [-0.22, 0.62]
Extraneous cognitive load	0.19 [-0.23, 0.61]	−0.08 [-0.50, 0.33]	**−0.30 [-0.72, 0.12]**
Positive affect	0.06 [-0.36, 0.48]	**0.32 [-0.09, 0.74]**	**0.25 [-0.17, 0.67]**
Negative affect	**0.51 [0.09, 0.94]**	**0.22 [-0.20, 0.63]**	**−0.32 [-0.74, 0.10]**

Note. For each row, effect size estimates are calculated based on follow-up pairwise t-tests (after running univariate ANCOVAs). Cohen's $d > |0.2|$ are depicted in bold. For a pairwise experimental comparison (*X* & *Y*), positive effect sizes denote evidence in favor of condition *X* and negative effect sizes denote evidence in favor of condition *Y*.

Table 5
Varimax-rotated components from principal component analysis.

	RC1 (Non-isomorphic conceptual understanding)	RC2 (Transfer)	RC3 (Isomorphic conceptual understanding)
Posttest Item 1 (financially wiser canton)	0.11	0.03	**0.99**
Posttest Item 2 (socialist ideology)	**0.82**	0.12	0.08
Posttest Item 3 (dataset anonymization)	**0.79**	−0.18	0.07
Posttest Item 4 (dataset normality)	−0.03	**0.98**	0.03

Note. For each dimension, loadings for representative posttest items are depicted in bold.

(item 4) and *isomorphic* conceptual understanding (item 1). Subsequently, we used estimates of the derived component scores as representative of these three posttest dimensions. Table 5 shows rotated loadings (eigenvectors) of the posttest items. Neither the MANCOVA nor univariate ANCOVAs (and Bonferroni-corrected follow-up t-tests) were significant for the posttest score and reasoning quality. Tables 3 and 4 summarize these results.

For the *isomorphic posttest*, students in the Failure-driven condition descriptively scored lower than students in the Productive Failure ($d = 0.31$) and Success-driven ($d = 0.50$) conditions. Effect size was low ($d < |0.2|$) when comparing the Productive Failure and Success-driven condition, suggesting similar posttest scores. BF_c corresponding to this overall observed trend was 6.08. However, when looking at *reasoning* across experimental conditions for the *isomorphic posttest*, we found a reversal in trend. Students in both the Failure-driven ($d = 0.40$) and Productive Failure ($d = 0.30$) conditions had a descriptively better quality of reasoning, relative to the Success-driven condition. There were no differences in reasoning quality between the Failure-driven and Productive Failure condition ($d < |0.2|$). BF_c corresponding to this overall observed trend in reasoning quality was 7.12.

For the *non-isomorphic posttest*, students in the Failure-driven condition scored similar to students in the Productive Failure and Success-driven conditions ($d < |0.2|$). The only posthoc pairwise comparison that showed $d > |0.2|$ was that between the Productive Failure and Success-driven condition, where the Success-driven condition students descriptively scored higher ($d = 0.34$). BF_c corresponding to this overall observed trend in posttest scores was 3.74. These odds were, however, 1.62 times weaker compared to trends in the isomorphic posttest scores. When looking at the *non-isomorphic posttest reasoning*, the only posthoc pairwise comparison that showed $d > |0.2|$ was that between the Failure-driven and Productive Failure condition. This comparison suggested that students in Failure-driven condition descriptively showed higher quality of reasoning ($d = 0.24$). Other pairwise differences between conditions suggested similar quality of reasoning ($d < |0.2|$). BF_c corresponding to this overall observed trend in reasoning quality was 3.09. These odds were, however, 2.3 times weaker compared to trends in the isomorphic posttest reasoning quality.

For the *transfer posttest*, we saw a reversal in trend for posttest scores (compared to isomorphic problem-solving). Here, students in the Failure-driven condition scored descriptively higher than students in both the Productive Failure ($d = 0.56$) and Success-driven ($d = 0.35$) conditions. Further, students in the Success-driven condition also scored descriptively higher than students in the Productive Failure ($d = 0.21$) condition, although the effect size was low. BF_c corresponding to this overall observed trend in posttest scores was 8.38. These odds were 1.38 and 2.24 times stronger compared to trends in the isomorphic and non-isomorphic posttest scores respectively. Similar to non-isomorphic problem-solving, the only pairwise comparison of *reasoning* quality for the *transfer posttest* that showed $d > |0.2|$ was that between the Failure-driven and Productive Failure condition – students in the Failure-driven condition descriptively showed higher quality of reasoning ($d = 0.21$). Other pairwise differences between conditions suggested similar quality of reasoning ($d < |0.2|$). BF_c corresponding to this overall observed trend in reasoning quality was 2.32. These odds were 3.06 and 1.33 times weaker compared to trends in the isomorphic and non-isomorphic posttest reasoning quality respectively.

Overall, these results *partially* support Hypothesis 1b regarding students in the Success-driven condition having the highest posttest performance (in terms of scores) for the isomorphic assessment (supported), lowest performance for the transfer assessment (*not* supported), and similar performance as the Failure-driven condition students on the non-isomorphic assessment (supported). When focusing on reasoning quality as an index of posttest performance, these results *do not* support hypothesis 1b.

5.2. Underlying mechanisms (RQ2)

The MANCOVA revealed significant differences in task experiences (assessed after the problem-solving phase) across conditions, Wilks' $\lambda = 0.78$, $F(14, 222) = 2.05$, $p = 0.015$. Univariate ANCOVAs showed significant differences in knowledge gap awareness, $F(2, 130) = 3.68$, $p = 0.028$, $\eta_p^2 = 0.06$ and marginally significant differences in negative affect, $F(2, 130) = 2.43$, $p = 0.093$, $\eta_p^2 = 0.04$. Tables 3 and 4 summarize these results.

Students in the Failure-driven condition reported descriptively higher knowledge gap awareness ($d = 0.53$), state curiosity ($d = 0.41$), and negative affect ($d = 0.51$), relative to the Productive Failure condition students. Other task experiences were similar between the two conditions ($d < |0.2|$). Failure-driven condition students also reported descriptively higher state curiosity ($d = 0.27$), affect (negative: $d = 0.22$, positive: $d = 0.32$), germane cognitive load ($d = 0.31$), and cognitive dissonance ($d = 0.27$), relative to the Success-driven condition. Extraneous cognitive load and knowledge gap awareness were similar between the Failure-driven and Success-driven conditions ($d < |0.2|$). Finally, we found that students in the Productive Failure condition reported descriptively higher cognitive dissonance ($d = 0.36$) and positive affect ($d = 0.25$), and lower knowledge gap awareness ($d = 0.52$), extraneous cognitive load ($d = 0.30$), and negative affect ($d = 0.32$), relative to the Success-driven condition. State curiosity and germane cognitive load were however similar between the two conditions ($d < |0.2|$).

Taken together, for cognitive dissonance and positive affect, these results support Hypothesis 2 regarding the higher intensity of triggered mechanisms in the Failure-driven and Productive Failure conditions, relative to students in the Success-driven condition. For state curiosity, negative affect and germane cognitive load, results only *partially* support Hypothesis 2, with only the Failure-driven condition students self-reporting higher scores than the Success-driven condition. Finally, for the remaining two underlying mechanisms of knowledge gap awareness and extraneous cognitive load, results *do not* support Hypothesis 2. The BF_c corresponding to the overall observed trends in students' task experiences for the problem-solving phase was usually in the similar positive/substantial range of 3–10 (5.82, 5.90, 8.38, 5.03, 4.37 and 5.12 for knowledge gap awareness, cognitive dissonance, state curiosity, germane cognitive load, extraneous cognitive load, and positive affect). Only the overall observed trend in negative affect had a relatively higher BF_c of 13.79.

Based on prior work (Loibl et al., 2017), we would expect these underlying preparatory mechanisms to differentially affect how students perceive follow-up instruction and consequently learn from it. On checking whether these differences manifested in the instruction exposure time, we found that Failure-driven condition students took

significantly longer time ($M = 316.37$, $SE = 17.32$) in part three of the lecture that discussed the canonical solution for the problem-solving task, $F(2, 130) = 5.2$, $p = 0.007$, $\eta_p^2 = 0.07$, relative to students in both the Productive Failure, $M = 245.51$, $SE = 17.72$, $d = 0.57$ and Success-driven, $M = 249.8$, $SE = 17.51$, $d = 0.57$ conditions.

5.3. Metacognitive calibration (RQ3)

The one-sample *t*-test showed that students in the *Failure-driven condition* were not well-calibrated in their initial problem-solving attempts, as reflected in the Δ (confidence, performance) being significantly different from 0 – they were under-confident for both attempt 1, $t(43) = -4$, $p < 0.001$, $d = -0.60$ and attempt 2, $t(39) = -2.94$, $p = 0.005$, $d = -0.46$. However, these students were well-calibrated during posttest, as reflected in Δ (confidence, performance) being not significantly different from 0 – isomorphic, $t(26) = 0.03$, $p = 0.975$, $d = 0.07$, non-isomorphic, $t(26) = -0.92$, $p = 0.364$, $d = -0.18$, and transfer, $t(26) = -0.84$, $p = 0.406$, $d = -0.16$ assessments.

Productive Failure condition students, on the other hand, showed an under-confidence bias for attempt 1 during the problem-solving phase, $t(42) = -4.14$, $p < 0.001$, $d = -0.63$. However, as these students kept generating more solutions during the second problem-solving attempt, they became well-calibrated, $t(42) = -1.17$, $p = 0.252$, $d = -0.20$. This effect was also seen during the isomorphic, $t(30) = -2.92$, $p = 0.772$, $d = -0.05$ and non-isomorphic, $t(30) = 0.93$, $p = 0.359$, $d = 0.17$ posttests. However, the major difference relative to the Failure-driven condition was seen during the transfer posttest, where these students expressed an over-confidence bias, $t(30) = 1.85$, $p = 0.075$, $d = 0.33$.

Finally, we found that students in the *Success-driven* condition were well-calibrated only for the transfer posttest, $t(35) = -0.31$, $p = 0.760$, $d = -0.05$. These students expressed an under-confidence bias during the initial problem-solving phase, both for attempt 1, $t(38) = -5.77$, $p < 0.001$, $d = -0.92$ and attempt 2, $t(32) = -2.35$, $p = 0.025$, $d = -0.41$. A similar under-confidence bias was also observed when these students solved the isomorphic, $t(35) = -2.28$, $p = 0.029$, $d = -0.38$ and the non-isomorphic, $t(35) = -1.76$, $p = 0.087$, $d = -0.29$ posttest. Table 6 summarizes these results.

Taken together, for the problem-solving phase, these results *do not* support Hypothesis 3 regarding better metacognitive calibration for students in the Failure-driven and Productive Failure conditions (relative to students in the Success-driven condition). For the posttest, however, the results *partially* support Hypothesis 3.

6. Discussion

6.1. Performance (RQ1)

During the *second problem-solving attempt*, students generated the maximum number of solutions in the Productive Failure condition and the least number of solutions in the Failure-driven condition. Not surprisingly, for the Productive Failure condition, this can be seen as a direct consequence of the task instructions that emphasized continuing solution generation in attempt 2 (similar to what students did in attempt 1). Note, however, that across both attempts, only one student in the Productive Failure condition generated a one-dimensional histogram (one of the three suboptimal scaffolds explicitly offered to the Failure-driven condition students). This means that despite a high number of generated solutions, students in the Productive Failure condition were not likely to naturally come up with suboptimal representations offered to the Failure-driven condition students. Further, the decreasing trend in solution quantity between the two scaffolding conditions can be attributed to the fact that in contrast to problematizing scaffolds, structuring scaffolds increase the certainty of action as they nudge students towards the canonical solution. This results in the generation of relatively more new solutions by incorporating information from these success-driven scaffolds.

For the *isomorphic posttest*, high(est) scores for students in the Success-driven condition might stem from them being exactly told the right way to proceed during the problem-solving phase (this was also subsequently asserted during the instruction phase). Not surprisingly, the initial relative dip in performance for the Failure-driven condition makes sense, because unlike other conditions, students exposed to problematizing scaffolds were focused on exploring the problem-space with suboptimal representations, instead of following an isolated (correct) solution pathway. However, the reversal in trend for reasoning quality demonstrates that higher performance (evidenced by highest posttest scores in the Success-driven condition) may not always be reflective of a high depth of understanding of the underlying domain principle (evidenced by lowest posttest reasoning quality).

For the *non-isomorphic posttest*, the Failure-driven condition was as powerful as the remaining experimental conditions, at least when looking at posttest scores. In contrast to trends for the isomorphic problem-solving posttest scores, this suggests that differences between experimental conditions start to disappear as a greater depth of understanding is required to tackle a problem situation. This understanding was empirically reflected in the highest reasoning quality demonstrated

Table 6
Metacognitive calibration (RQ3) across experimental conditions.

	Calibration bias: Δ (Confidence, Performance)			Cohen's *d* [95% CI]
	Under-confident	Over-confident	Well-calibrated	
Failure-driven				
Problem-solving phase attempt 1	✓			−0.60 [-0.92, −0.28]
Problem-solving phase attempt 2	✓			−0.46 [-0.79, −0.14]
Isomorphic posttest			✓	0.07 [-0.37, 0.38]
Non-isomorphic posttest			✓	−0.18 [-0.56, 0.20]
Transfer posttest			✓	−0.16 [-0.54, 0.22]
Productive Failure				
Problem-solving phase attempt 1	✓			−0.63 [-0.96, −0.30]
Problem-solving phase attempt 2			✓	−0.20 [-0.55, 0.14]
Isomorphic posttest			✓	−0.05 [-0.40, 0.30]
Non-isomorphic posttest			✓	0.17 [-0.19, 0.52]
Transfer posttest		✓		0.33 [-0.03, 0.69]
Success-driven				
Problem-solving phase attempt 1	✓			−0.92 [-1.93, −0.54]
Problem-solving phase attempt 2	✓			−0.41 [-0.76, −0.05]
Isomorphic posttest	✓			−0.38 [-0.72, −0.04]
Non-isomorphic posttest	✓			−0.29 [-0.62, 0.04]
Transfer posttest			✓	−0.05 [-0.38, 0.28]

Note. Checkmarks are based on significant results from the one-sample *t*-test ($p < 0.05$). For each row, positive effect sizes denote over-confidence bias, negative effect sizes denote under-confidence bias and effect sizes close to 0 denote well-calibrated judgment of performance.

by students in the Failure-driven condition. Further, we also found the Success-driven condition to be relatively more efficacious compared to the Productive Failure condition in terms of posttest scores. What might explain this advantage? Because students have opportunities to fail during the first problem-solving attempt and receive scaffolding towards the canonical solution during the second attempt, the Success-driven condition can alternatively be perceived as exposing students to a smaller iteration (cycle) of Productive Failure even before formal instruction happens. Support during the second problem-solving attempt and follow-up instruction present redundant scaffolding opportunities (Tabak, 2004) for understanding different conceptual task elements. This might, in turn, have resulted in better learning for students in the Success-driven condition (relative to the Productive Failure condition).

For the *transfer posttest*, a reversal in trend for posttest scores compared to isomorphic problem-solving suggests that the preparatory benefits of explicit problematizing start becoming salient when posttest problem situations require flexible adaptation of what is learned from instruction and/or the ability to re-learn. For students in the Success-driven condition, high accessibility of correct information from the scaffolds presented in the problem-solving phase (that can be posited to help in isomorphic problem-solving) does not reliably reflect their ability to transfer. We further found that students in the Failure-driven condition showed the highest reasoning quality for the transfer posttest (similar to the trends for non-isomorphic problem-solving). This indicates that not only did these students perform better compared to the other conditions, but they were also relatively more fluent in explaining their solution rationale.

6.2. Underlying mechanisms (RQ2)

Comparison of students in the Failure-driven and Productive Failure conditions suggests that explicit problematizing in the Failure-driven condition facilitates (to a greater extent) the realization of what is known and not known about the targeted concept, as well as students' desire to know more about the canonical solution to fill these knowledge gaps. Further, a higher lecture viewing time for the part of the follow-up instruction discussing the canonical solution might afford additional reflection and comparison opportunities for students in the Failure-driven condition, in turn allowing them to revise their understanding of the targeted concept. Students learn more deeply when they are able to explain and think about the inter-connections between new and existing knowledge, than when they do not (Chi, 2009).

The relatively higher negative affect in the Failure-driven condition might be attributed to the *explicitly* induced opportunities for failure in the problem-solving process. However, counter to studies that have found negative emotions during learning to result in narrowed attentional focus (Kaspar & König, 2012), longer times required to reach mastery levels and lower performance on transfer tasks (Brand, Reimer, & Opwis, 2007; Pekrun & Linnenbrink-Garcia, 2012), we found that students in the Failure-driven condition consistently exhibited the highest reasoning quality. In fact, these students even outperformed students in the Productive Failure (as well as the Success-driven) condition on the transfer posttest. The relatively more careful and analytical, detailed, and rigid manners of processing information (Knörzer, Brünken, & Park, 2016) might have attributed to the facilitating effect of negative emotional experiences during the problem-solving phase. Thus, explicit failure-driven problem-solving experiences that evoke negative affect may not always be undesirable for learning (Sinha, 2021).

A comparison of the Failure-driven and Success-driven conditions suggests that the relatively higher levels of experienced dissonance in the Failure-driven condition might be attributed to students spending effort in enduring an uncomfortable task situation whose outcome is not readily verifiable. Problematizing scaffolds challenge students' current understanding, force reasoning with suboptimal representations, and provide no straightforward way to move towards the canonical solution. In contrast to the Success-driven condition, this is likely to result in problem-solving behaviors featuring relatively greater trial and error characteristics that build on situational feedback from the problem-solving environment, in order to reduce dissonance. In other words, the Failure-driven condition creates a legitimate need for additional sensemaking activities compared to the Success-driven condition. Despite the potential for high discrepancy with the canonical solution, we posit that this additional learning stemming from sensemaking opportunities with the problematizing scaffolds holds high preparatory benefits. Although we did not find direct evidence for this conjecture in the immediate problem-solving performance (both Failure-driven and Success-driven conditions had similar solution quality during attempt 2), we saw delayed benefits in posttest reasoning quality and transfer outcomes. This clearly reflects the incommensurability between performance and learning (Kapur, 2016; Soderstrom & Bjork, 2015).

Finally, we also found that getting exposed to new learning materials via the multi-step scaffold presentation (irrespective of their type and whether the presented ideas are optimal or suboptimal) induces similar levels of knowledge gap awareness. In contrast to the Productive Failure condition where students work with their own ideas (what they know), the scaffolding conditions provide relatively greater opportunities for raising awareness of what is not known.

6.3. Metacognitive calibration (RQ3)

The bias towards under-confidence during the problem-solving phase for all conditions is plausible. Our study design exposes students to novel problem solutions and asks them to generate and explain the rationale for multiple representations and solutions. Indeed, this is not a common norm in problem-solving practice. However, as students got more conscious about their acquired knowledge after follow-up instruction, unexpectedly, they did not estimate their performance in posttests more accurately. We observed salient differences in metacognitive calibration as students worked on the posttests.

The fact that students in the Failure-driven condition were metacognitively well-calibrated in the posttest might explain why their relative efficacy increased as they moved on from the isomorphic to the non-isomorphic to the transfer questions. On the other hand, surprisingly, students in the Productive Failure condition who were metacognitively well-calibrated for the isomorphic and non-isomorphic posttests, exhibited an over-confidence bias in the transfer posttest. Relying on biased evaluations can harm performance by misleading effort regulation (Ackerman & Thompson, 2017), for example by prematurely stopping consideration of other (more relevant) problem-solving strategies. This might have worsened the performance of students exposed to the Productive Failure condition on the transfer posttest. Finally, for students in the Success-driven condition, being mostly under-confident suggests that although they followed the presented scaffolds to move closer to the canonical solution, they may not have necessarily understood the underlying concept/idea.

7. General discussion and conclusion

In the current study, we compared three variants of the PS-I design where students received explicit problem-solving scaffolds that either nudged them towards optimal solutions (Success-driven), suboptimal solutions (Failure-driven) or received no explicit scaffolds in the problem-solving phase (Productive Failure). Overall, our results are consistent with results reported in Sinha et al. (2020) and hold up under a controlled study as well. We found an overall efficacy of failure-driven preparatory activities over success-driven activities on (a) non-isomorphic conceptual understanding (*similar* posttest scores, but *higher* reasoning quality in the Failure-driven condition) and additionally, (b) transfer outcomes (*higher* posttest scores, *higher* reasoning quality). We further found that students in the Productive Failure condition had relatively worse scores and reasoning quality for non-isomorphic conceptual understanding and transfer outcomes,

compared to students exposed to failure-driven scaffolding. The only posttest dimension for which students in the Success-driven and Productive Failure conditions outperformed the Failure-driven condition was that of isomorphic conceptual understanding.

One relevant design-level difference with respect to prior experimental studies comparing scaffolding (typically, success-driven) and an unscaffolded (pure Productive Failure) condition, for example, Kapur (2011) and Loibl and Rummel (2014a), is that we had a two-phase problem-solving design. In contrast to earlier single-phase implementations in the literature with scaffolding integrated throughout the problem-solving phase, students in the Failure-driven and Success-driven conditions here went through an initial attempt at generating representation and solution methods in the absence of any scaffolds. Differential scaffolds were only subsequently introduced in a second phase, affording students opportunities to generate more representations and/or revise representations from the first phase. Thus, our learning design struck a balance between first giving students complete agency and allowing them to freely invent their own representations for learning, and only then, externally providing plausible/intelligible representations to construct. This might explain the efficacy of failure-driven and success-driven scaffolding in learning from instruction, relative to the unscaffolded condition (that only emphasized invention).

When compared to the average effect size disfavoring PS-I (Hedge's g −0.08 [95% CI -0.34, 0.28]) relative to scaffolded PS-I learning designs for a similar postgraduate student population and learning outcomes such as ours (Sinha & Kapur, 2021b), effect sizes for the current study exceed or fall closer to the lower end of the confidence interval, at least when it comes the impact of explicit failure-driven scaffolding on non-isomorphic conceptual understanding and transfer. In that light, our results bear tremendous practical significance (Hill et al., 2008). Our results also expand the explanatory basis for why the reported trends might hold, in particular, illuminating not just cognitive mechanisms such as knowledge gap awareness and cognitive load, but also students' perceived positive and negative affect, and metacognitive mechanisms like calibration bias. The inclusion of such a wide range of underlying mechanisms for comparative evaluation of different preparatory problem-solving approaches makes a strong case for wide-scale adoption of the developed materials.

It is important to highlight, however, that the reported results are correlational. We could not test directly for causality (and/or temporality) because of lacking a large enough sample size and a longitudinal design with repeated measurement of mediators (learning mechanisms).[11] Also, methodologically, because the specific hypotheses we tested using Bayesian informative hypotheses evaluation ANCOVA corresponded to the observed trends in marginal means for the dependent variables across experimental conditions (for which we obtained moderate effect sizes, despite non-significance), it was plausible that the BF_c revealed moderate evidence for several comparisons.[12] One caveat in using Bayesian informative hypotheses evaluation ANCOVA, however, is that there might exist other unconsidered hypotheses for which the support in the data might be larger.

For the design of scaffolding in learning through problem-solving, our results regarding the pedagogical benefits of explicit failure-driven scaffolding imply their inclusion to serve as an effective preparatory activity in classroom teaching practices. Taken together, the findings from the quasi-experimental study (Sinha et al., 2020) and our controlled study support the deliberate design of failure-driven experiences before formal instruction, if educators wish to foster students' ability to take knowledge learned in one context and apply it in novel contexts. More generally, educators might consider ways in which students, via explicit and deliberately-designed failures in preparatory problem-solving, can become aware of the limitations of their prior knowledge. Inviting initial failure-driven participation by creating in-situ opportunities for suboptimal representation generation might provide a strong foundation for reorganization of existing knowledge when students are exposed to instruction. In contrast to the uncertainty-reducing nature of success-driven scaffolding, failure-driven scaffolding can be conceived as a route to harnessing the potential benefits of immersing students in uncertainty (Metz, 2004). Uncertain situations demand "searching, hunting, inquiring, to find material that will resolve the doubt, settle and dispose of the perplexity" (Dewey, 1933, p. 121). The scope of learning with failure-driven scaffolding is therefore high.

Research that delves more deeply into other forms of failure-driven scaffolding could also be a promising future direction. For example, explicitly presenting students with datasets or situations where their current solution does not apply, might signal a need to gather more evidence to integrate with prior evidence. The selection of such contrasting cases as scaffolds on-the-fly, in order to question the optimality of an already developed solution approach, is an exciting research direction. More broadly, this ties into the idea of adaptive scaffold presentation (Aleven et al., 2017; Roll, 2009). Such personalization, which can iteratively and naturally gauge students' understanding (e.g., by holding casual conversations) and deliver scaffolds out of order (as and when necessary), is likely to improve students' metacognition about what they know and what they do not know, and in turn, lead to improved learning from instruction.

We would like to emphasize, however, that the adoption of *deliberate, guided failure* as a scaffolding strategy during preparatory problem-solving should go hand-in-hand with fostering positive teacher-student relationships (Jennings & Greenberg, 2009). For instance, students are more likely to trust the pedagogical value of classroom activities, if they have a good interpersonal rapport with the teacher, and trust the teacher to show genuine interest in (and meet) their developmental, emotional, and academic needs. For teachers, therefore, it is equally critical to cultivate positive relationships in the classroom as well as adopt a pedagogical value-style framing of failure-driven scaffolding activities (that clearly emphasizes the utility of engaging in such activities), in order to make students increasingly more comfortable with the uncomfortable.

Author note

Tanmay Sinha and Manu Kapur, Professorship for Learning Sciences and Higher Education, ETH Zürich, Switzerland.

We thank Stefan Wehrli for providing us with the Decision Sciences Lab infrastructure (ETH Zürich) for running the study, and making sure that the study sessions run smoothly. We thank Cornelia Schnyder (University of Zürich) for assisting us with recruitment of study participants. We also acknowledge the efforts of Giordano Giannoccolo from the Decision Sciences Lab for helping automate the study workflow and for handling participant reimbursement. Thanks to Maya Spannagel (University of Zürich) for assistance in running the study sessions. Thanks to Anne Deiglmayr (University of Leipzig) and Elsbeth Stern (ETH Zürich) for sharing the calculus concept inventory used in the study. Thanks to colleagues from the Professorship for Learning Sciences and Higher Education, ETH Zürich for helpful feedback regarding this

[11] Concurrent mediation analysis comparing the Failure-driven and Productive Failure conditions, however, showed knowledge gap awareness to significantly mediate the impact on the transfer posttest, $\beta = 0.18$ ($SE = 0.11$), 95% bias-corrected percentile bootstrap CI [0.01, 0.46] (see supplementary materials for details).

[12] Bayesian error probabilities (uncertainty about a Hypothesis H) are computed conditional on the information in the observed data. The Bayesian error associated with a preference of that hypothesis will therefore be smaller (or, P (H | data) will be higher) for a data set with a high Cohen's d (e.g., 0.4) than for a data set with a low Cohen's d (e.g., 0.1). In other words, if a null hypothesis specifying the equality of experimental conditions is true, it is much less likely to observe a Cohen's d of 0.4 than a Cohen's d of 0.1 (and vice-versa).

work. A final note of thanks to anonymous reviewers whose mentorship helped improve the quality of this manuscript.

CRediT authorship contribution statement

Tanmay Sinha: Conceptualization, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **Manu Kapur:** Conceptualization, Writing – review & editing, Supervision.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.learninstruc.2021.101488.

References

Ackerman, R., & Thompson, V. A. (2017). Meta-Reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences, 21*(8), 607–617. https://doi.org/10.1016/j.tics.2017.05.004

Aleven, V., Connolly, H., Popescu, O., Marks, J., Lamnina, M., & Chase, C. (2017). An adaptive coach for invention activities. In *International Conference on Artificial Intelligence in Education* (pp. 3–14). https://doi.org/10.1007/978-3-319-61425-0_1

Aleven, V., McLaughlin, E. A., Glenn, R. A., & Koedinger, K. R. (2016). Instruction based on adaptive learning technologies. *Handbook of Research on Learning And Instruction*, 522–560.

Aleven, V., Stahl, E., Schworm, S., Fischer, F., & Wallace, R. (2003). Help seeking and help design in interactive learning environments. *Review of Educational Research, 73*(3), 277–320. https://doi.org/10.3102/00346543073003277

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64*, 417–444. https://doi.org/10.1146/annurev-psych-113011-143823

Brand, S., Reimer, T., & Opwis, K. (2007). How do we learn in a negative mood? Effects of a negative mood on transfer and learning. *Learning and Instruction, 17*(1), 1–16. https://doi.org/10.1016/j.learninstruc.2006.11.002

Button, S. B., Mathieu, J. E., & Zajac, D. M. (1996). Goal orientation in organizational research: A conceptual and empirical foundation. *Organizational Behavior and Human Decision Processes, 67*, 26–48. https://doi.org/10.1006/obhd.1996.0063

Chi, M. T. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science, 1*(1), 73–105. https://doi.org/10.1111/j.1756-8765.2008.01005.x

Clifford, M. M. (1984). Thoughts on a theory of constructive failure. *Educational Psychologist, 19*(2), 108–120. https://doi.org/10.1080/00461528409529286

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7–29. https://doi.org/10.1177/0956797613504966

De Jong, T., Sotiriou, S., & Gillet, D. (2014). Innovations in STEM education: The go-lab federation of online labs. *Smart Learning Environments, 1*(1), 1–16. https://doi.org/10.1186/s40561-014-0003-6

Dewey, J. (1933). *How we think: A restatement of the relation of reflective thinking to the educative process* (Vol. 8). Lexington, MA: Heath.

Dweck, C. S. (1992). Article commentary: The study of goals in psychology. *Psychological Science, 3*, 165–167. https://doi.org/10.1111/j.1467-9280.1992.tb00019.x

Epstein, J. (2007). Development and validation of the calculus concept inventory. In *Proceedings of the ninth international conference on mathematics education in a global community* (pp. 165–170).

Festinger, L. (1962). *A theory of cognitive dissonance*. Palo Alto, CA: Stanford University Press.

Fielstein, E., Klein, M. S., Fischer, M., Hanan, C., Koburger, P., Schneider, M. J., et al. (1985). Self-esteem and causal attributions for success and failure in children. *Cognitive Therapy and Research, 9*, 381–398. https://doi.org/10.1007/BF01173088

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

Glogger-Frey, I., Gaus, K., & Renkl, A. (2017). Learning from direct instruction: Best prepared by several self-regulated or guided invention activities? *Learning and Instruction, 51*, 26–35. https://doi.org/10.1016/j.learninstruc.2016.11.002

Hammer, D. (2000). Student resources for learning introductory physics. *American Journal of Physics, 68*, S52–S59. https://doi.org/10.1119/1.19520

Harmon-Jones, E., Price, T. F., Gable, P. A., & Peterson, C. K. (2014). Approach motivation and its relationship to positive and negative emotions. In M. M. Tugade, M. N. Shiota, & L. D. Kirby (Eds.), *Handbook of positive emotions* (pp. 103–118). Guilford Press.

Harter, S. (2012). *Self-perception profile for adolescents: Manual and questionnaires*. Univeristy of Denver, Department of Psychology.

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.17508606.2008.00061.x

Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: A response to kirshner, sweller, and Clark (2006). *Educational Psychologist, 42*, 99–107. https://doi.org/10.1080/00461520701263368

Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods, 24*(5), 539. https://doi.org/10.1037/met0000201

Holmes, N. G., Day, J., Park, A. H., Bonn, D. A., & Roll, I. (2014). Making the failure more productive: Scaffolding the invention process to improve inquiry behaviors and outcomes in invention activities. *Instructional Science, 42*(4), 523–538. https://doi.org/10.1007/s11251-013-9300-7

Izard, C. E. (1977). Anger, disgust, and contempt and their relationship to hostility and aggression. In C. Izard (Ed.), *Human emotions* (pp. 329–354). Springer. https://doi.org/10.1007/978-1-4899-2209-0_13.

Jennings, P. A., & Greenberg, M. T. (2009). The prosocial classroom: Teacher social and emotional competence in relation to student and classroom outcomes. *Review of Educational Research, 79*(1), 491–525. https://doi.org/10.3102/0034654308325693

Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(6), 1420. https://doi.org/10.1037/0278-7393.20.6.1420

Kapur, M. (2011). A further study of productive failure in mathematical problem solving: Unpacking the design components. *Instructional Science, 39*(4), 561–579. https://doi.org/10.1007/s11251-010-9144-3

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

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. *The Journal of the Learning Sciences, 21*(1), 45–83. https://doi.org/10.1080/10508406.2011.591717

Kapur, M., & Kinzer, C. K. (2009). Productive failure in cscl groups. *International Journal of Computer-Supported Collaborative Learning, 4*(1), 21–46. https://doi.org/10.1080/00461520.2016.1155457

Kaspar, K., & König, P. (2012). Emotions and personality traits as high-level factors in visual attention: A review. *Frontiers in Human Neuroscience, 6*, 321. https://doi.org/10.3389/fnhum.2012.00321

Knol, M. H., Dolan, C. V., Mellenbergh, G. J., & van der Maas, H. L. (2016). Measuring the quality of university lectures: Development and validation of the instructional skills questionnaire (ISQ). *PloS One, 11*(2), Article e0149163. https://doi.org/10.1371/journal.pone.0149163

Knörzer, L., Brünken, R., & Park, B. (2016). Facilitators or suppressors: Effects of experimentally induced emotions on multimedia learning. *Learning and Instruction, 44*, 97–107. https://doi.org/10.1016/j.learninstruc.2016.04.002

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121.

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*, Article 101785. https://doi.org/10.1016/j.cedpsych.2019.101785

Lee, H. S., & Anderson, J. R. (2013). Student learning: What has instruction got to do with it? *Annual Review of Psychology, 64*, 445–469. https://doi.org/10.1146/annurev-psych-113011-143833

Leighton, J. P., Tang, W., & Guo, Q. (2015). Developing and validating the attitudes towards mistakes inventory (atmi): A self-report measure. In *Proceedings of the annual Meeting of the national Council on Measurement in education*.

Leppink, J., Paas, F., Van Gog, T., van Der Vleuten, C. P., & Van Merrienboer, J. J. (2014). Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learning and Instruction, 30*, 32–42. https://doi.org/10.1016/j.learninstruc.2013.12.001

Levin, D. T., Harriott, C., Paul, N. A., Zhang, T., & Adams, J. A. (2013). Cognitive dissonance as a measure of reactions to human-robot interaction. *Journal of Human-Robot Interaction, 2*(3), 3–17. https://doi.org/10.5898/JHRI.2.3 (Levin).

Likourezos, V., & Kalyuga, S. (2017). Instruction-first and problem-solving-first approaches: Alternative pathways to learning complex tasks. *Instructional Science, 45*(2), 195–219. https://doi.org/10.1007/s11251-016-9399-4

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). 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/s11251-0139282-5

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

Marei, H. F., Donkers, J., Al-Eraky, M. M., & Van Merrienboer, J. J. (2019). Collaborative use of virtual patients after a lecture enhances learning with minimal investment of cognitive load. *Medical Teacher, 41*(3), 332–339. https://doi.org/10.1080/0142159X.2018.1472372

Metcalfe, J. (2017). Learning from errors. *Annual Review of Psychology, 68*, 465–489. https://doi.org/10.1146/annurev-psych-010416-044022

Metz, K. E. (2004). Children's understanding of scientific inquiry: Their conceptualization of uncertainty in investigations of their own design. *Cognition and Instruction, 22*(2), 219–290. https://doi.org/10.1207/s1532690xci2202_3

Mikulincer, M. (1994). Human learned helplessness: A coping perspective. In L. R. Huesmann (Ed.), *The plenum series in social/clinical psychology*. New York, NY: Plenum Press.

Naylor, F. D. (1981). A state-trait curiosity inventory. *Australian Psychologist, 16*, 172–183. https://doi.org/10.1080/00050068108255893

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

Pekrun, R., & Linnenbrink-Garcia, L. (2012). Academic emotions and student engagement. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 259–282). Berlin, Germany: Springer. https://doi.org/10.1007/978-1-4614-2018-7_12.

Pintrich, P. R. (1991). *A manual for the use of the motivated strategies for learning questionnaire (MSLQ). (Report NCRIPTAL-91-B-004). Office of educational Research and improvement*. ERIC Document Reproduction Service No. ED338122.

Quintana, C., Reiser, B. J., Davis, E. A., Krajcik, J., Fretz, E., Duncan, R. G., et al. (2004). A scaffolding design framework for software to support science inquiry. *The Journal of the Learning Sciences, 13*, 337–386. https://doi.org/10.1207/s15327809jls1303_4

Reiser, B. J. (2004). Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *The Journal of the Learning Sciences, 13*(3), 273–304. https://doi.org/10.1207/s15327809jls1303_2

Roebers, C. M. (2017). Executive function and metacognition: Towards a unifying framework of cognitive self-regulation. *Developmental Review, 45*, 31–51. https://doi.org/10.1016/j.dr.2017.04.001

Roll, I. (2009). *Structured invention Tasks to prepare Students for future learning: Means, mechanisms, and cognitive processes (doctoral dissertation)*. Pittsburgh: Carnegie Mellon University.

Schwartz, D. L., Catherine, C. C., & Bransford, J. D. (2012). Resisting overzealous transfer: Coordinating previously successful routines with needs for new learning. *Educational Psychologist, 47*(3), 204–214. https://doi.org/10.1080/00461520.2012.696317

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

Silvia, P. J. (2009). Looking past pleasure: Anger, confusion, disgust, pride, surprise, and other unusual aesthetic emotions. *Psychology of Aesthetics, Creativity, and the Arts, 3*(1), 48. https://doi.org/10.1037/a0014632

Sinha, T., Kapur, M., West, R., Catasta, M., Hauswirth, M., & Trninic, D. (2020). Differential benefits of explicit failure-driven and success-driven scaffolding in problem-solving prior to instruction. *Journal of Educational Psychology*. Advance online publication. doi: 10.1037/edu0000483.

Sinha, T., & Kapur, M. (2021a). *When problem-solving followed by instruction works: Evidence for productive failure*. Manuscript under review.

Sinha, T., & Kapur, M. (2021b). *From problem-solving to sensemaking: A comparative meta-analysis of preparatory approaches for future learning*. Manuscript under review.

Sinha, T. (2021). *Enriching problem-solving followed by instruction with explanatory accounts of emotions*. Manuscript under review.

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science, 10*(2), 176–199. https://doi.org/10.1177/1745691615569000

von Soest, T., Wichstrom, L., & Kvalem, I. L. (2016). The development of global and domain-specific self-esteem from age 13 to 31. *Journal of Personality and Social Psychology, 110*, 592–608. https://doi.org/10.1037/pspp0000060

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*(3), 271.

Tabak, I. (2004). Synergy: A complement to emerging patterns of distributed scaffolding. *The Journal of the Learning Sciences, 13*(3), 305–335. https://doi.org/10.1207/s15327809jls1303_3

Tulis, M., & Ainley, M. (2011). Interest, enjoyment and pride after failure experiences? Predictors of students' state-emotions after success and failure during learning in mathematics. *Educational Psychology, 31*(7), 779–807. https://doi.org/10.1080/01443410.2011.608524

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197–221. https://doi.org/10.1080/00461520.2011.611369

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The panas scales. *Journal of Personality and Social Psychology, 54*, 1063. https://doi.org/10.1037/0022-3514.54.6.1063

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17*, 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x