Impact of Explicit Failure and Success-driven Preparatory Activities on Learning

Tanmay Sinha[1], **Manu Kapur**[1], **Robert West**[2], **Michele Catasta**[3], **Matthias Hauswirth**[4], **Dragan Trninic**[1]
[1] ETH Zürich, Switzerland, [2] EPFL Lausanne, Switzerland, [3] Stanford University, USA, [4] University of Lugano, Switzerland

Abstract

Unscaffolded problem-solving before receiving instruction can give students opportunities to entertain their exploratory hypotheses at the expense of experiencing initial failures. Prior literature has argued for the efficacy of such Productive Failure (PF) activities in preparing students to "see" like an expert. Despite growing understanding of the socio-cognitive mechanisms that affect learning from PF, the necessity of success or failure in initial problem-solving attempts is still unclear. Consequently, we do not know yet whether some ways of succeeding or failing are more efficacious than others. Here, we report empirical evidence from a recently concluded classroom PF intervention (N=221), where we designed scaffolds to explicitly push student problem-solving towards success via structuring, but also radically, towards failure via problematizing. Our rationale for explicit failure scaffolding was rooted in facilitating problem-space exploration. We subsequently compared the differential preparatory effects of success-driven and failure-driven problem-solving on learning from subsequent instruction. Results suggested explicit failure scaffolding during initial problem-solving to have a higher impact on conceptual understanding, compared to explicit success scaffolding. This trend was more salient for the task topic with greater difficulty.

Keywords: Classroom Study; Productive Failure; Scaffolding

Introduction

Substantial research has demonstrated the efficacy of learning approaches where problem-solving as a preparatory activity precedes instruction (PS-I). PS-I includes (i) an initial problem-solving phase where students explore solutions to complex problems based on concepts they haven't formally learnt yet, and (ii) a subsequent explicit instruction phase where a coach introduces formalisms of the targeted concepts along with the canonical solution. Research suggests that PS-I is an effective learning design that improves student's conceptual understanding and positively impacts how well they transfer their knowledge to novel problem-solving contexts (Loibl, Roll, & Rummel, 2017).

A particular variant of the PS-I design that embodies learning from failure is Productive Failure (PF) (Kapur & Bielaczyc, 2012). PF comprises rich problem design that affords multiple representations and solution methods (RSMs), and follow-up instruction that compares and contrasts student-generated solutions with the canonical one. The positive benefits of approaches implemented based on the PS-I design (e.g., PF, Invent with Contrasting Cases (Schwartz & Martin, 2004)) have been attributed to different cognitive mechanisms. These include intentional activation of relevant prior knowledge, enhancement of students' awareness of the problem situation and own knowledge gaps, focused attention on search for deeper patterns rather than surface characteristics, and effortful retrieval to resolve incongruity. Some posited socio-emotional mechanisms include increased motivation to learn targeted concepts and elicitation of curiosity (Kapur & Bielaczyc, 2012; Loibl et al., 2017).

Research Gap

Despite PS-I designs often working better compared to traditional instructional approaches (usually direct instruction) on the acquisition of conceptual knowledge and/or transfer, there is a considerable variation in effect sizes (Cohen's d = 1.12$\pm$ 0.54) (Loibl et al., 2017). This has spurred lines of inquiry into systematically analyzing reasons for failure of PS-I approaches (Sinha & Kapur, 2019), and developing ways to improve overall effectiveness of the learning design. One prominent area of focus has been the initial problem-solving phase. Here, research has started to investigate the impact of scaffolding student solutions on fostering conceptually sound and transferable learning (Kapur, 2011; Loibl & Rummel, 2014). Despite growing research in the PS-I design space, we don't have conclusive evidence yet.

Templates of successful problem-solving usually aim at pro-active error elimination, and directing student's attention to the task by providing immediate feedback. Such instruction has the advantage of helping students perform the correct procedure. However, this may not always imply that students engage in optimal reasoning or acquire high depth of understanding of domain principles. Evidence favoring success-driven (SD) learning in PS-I suggests the presence of an association between successful problem-solving during the problem-solving phase and learning from instruction (e.g., Chin, Chi, and Schwartz (2016); Schwartz, Chase, Oppezzo, and Chin (2011); Loibl and Rummel (2014); Schalk, Schumacher, Barth, and Stern (2017); Chase and Klahr (2017)). However, attempts to scaffold such success, both cognitively (e.g., Kapur (2011); Loibl and Rummel (2014)) and metacognitively (e.g., Holmes, Day, Park, Bonn, and Roll (2014); Roll et al. (2018)), have been largely unsuccessful.

Templates of exploratory or unsuccessful problem-solving, on the other hand, hold the view that acquisition of solution schema is not the solitary goal of learning through problem-solving (Schwartz & Martin, 2004; Kapur & Bielaczyc, 2012). It is equally important to develop the cognitive and socio-emotional prerequisites to prepare novice students to see like an expert. Therefore, one should provide opportunities that help students develop awareness and appreciation for what is known and not known. Instructional attempts that increase chances of failure during problem-solving have the advantage of stimulating student's initiative in gaining knowledge. However, students may not spontaneously come back to the right track if an incorrect problem representation is invoked and they continue to work on it. Evidence disfavoring SD learning in PS-I suggests that a lack of success when the problem-solving phase is implicitly scaffolded (e.g. Aleven et al. (2017); Roelle and Berthold (2016); Mazziotti, Rummel, and Deiglmayr (2016)) or left unscaffolded (e.g., Kapur

and Bielaczyc (2012)) does not harm learning. Providing no explicit cognitive or metacognitive support is imperative in view of giving students complete agency in solution generation. A consequential side-effect is that the likelihood of experiencing failures increases.

However, there is no PS-I research that looks at explicitly scaffolding problem-solving phase towards failure. This sets up the guiding question of whether and to what extent is success or failure during initial problem-solving necessary for learning from PS-I. How does increasing likelihood of students experiencing success or failure differentially prepare them to learn from the instruction at a deeper conceptual level? Are some ways of succeeding or failing more efficacious than others? To answer these questions, we design SD or failure-driven (FD) scaffolds for the problem-solving phase, as inputs into a classroom PS-I intervention. Evidence for impact of these scaffolds on learning from PF is discussed.

Method

Participants and Task Domain

We conducted a classroom PS-I intervention with N=221 students in an introductory data science course offered at a large public university in Switzerland. Based on data from a previous course iteration, two topics Spurious Correlation (SC) and Anscombe's Quartet (AQ) were chosen to develop learning materials. Problem-solving based on these topics had demonstrated different initial failure rates, and different levels of improvement after students were presented with clues pointing them to the correct answer (SC task, 40% $\rightarrow$ 23%; AQ task, 81% $\rightarrow$ 38%). The SC learning goal was to help students tease apart the difference between strong versus meaningful relationships among dataset variables. The AQ learning goal was to help students understand the complementary importance of numerical and graphical representations in reasoning with data. Students worked individually in an online problem-solving environment (Python Jupyter notebook) that was dynamically executable, and helped in offloading procedural or syntactical aspects of the computation required (for task details, see www.tinyurl.com/CogSci2019Tasks).

Experimental Design and Scaffolding Rationale

A mixed experimental design was followed. Scaffolding in initial problem-solving (SD, FD) was the between-subject variable, and problem-solving topic (SC, AQ) was the within-subject variable. Students were randomly assigned to experimental conditions, and ordering of problem-solving topics was counterbalanced within each condition. We had two conditions representative of SD scaffolding with varying degrees of specificity, and two conditions representative of FD scaffolding with varying levels of suboptimality. For all four conditions, the instruction phase was kept constant. Student solutions were compared and contrasted with the canonical one.

The rationale for the concrete design of scaffolding in our research was inspired by mechanisms of structuring and problematizing student work (Reiser, 2004). Structuring scaffolds reduce degrees of freedom to lower task complexity, help students maintain direction, and make problem-solving tractable. Problematizing scaffolds increase degrees of freedom to challenge student's current understanding, and highlight discrepancies between what they might generate and critical/canonical task features. We chose an initial set of structuring and problematizing scaffolds in line with keeping the generative characteristics of the problem-solving phase intact, as well as explicitly increasing success or failure likelihood as the intended design rationale.

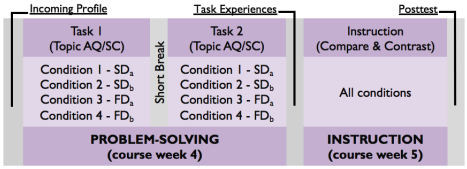

Figure 1: Experimental Design. SD_a, SD_b, FD_a, FD_b are two instantiations of success-driven (SD) and failure-driven (FD) scaffolding in the problem-solving phase respectively.

Structuring scaffolds included a combination of prompts, hints and bottom-out hints for different task topics. Prompts point students to the problem conditions that should likely remind them of the knowledge component's relevance. Very little information is divulged, thus encouraging students to do most of the thinking themselves. Our design of hints incorporates the idea of teaching students the knowledge component that is actually relevant in the current problem-solving context (what to do but not how). Finally, bottom-out hints tell students precise (and potentially optimal) ways of moving ahead in the problem-solving task. Such a scaffolding sequence mimics the behavior of expert human tutors, and is almost universally used in tutoring systems (VanLehn, 2011).

Problematizing scaffolds included asking students to explicitly generate suboptimal RSMs to facilitate problem-space exploration in a more comprehensive manner, rather than following an isolated solution path. In essence, students are led towards questionable decision-making by being asked to consider a subset of conceptual domain factors (that don't lead to the canonical solution), and reason with those partially-gained insights. No former PS-I work has looked directly into such "explicit" failure scaffolding. However, one could view the classic PF design as providing "implicit" opportunities for students to create suboptimal RSMs (Kapur & Bielaczyc, 2012). This is because rich problem design "inherently" affords multiple RSM generation, and targets concepts students haven't learnt yet. Work on preparatory benefits of vicarious failure activities before receiving instruction suggests the evaluation of suboptimal or failed RSMs generated by others as a significant predictor of learning (Kapur, 2014).

As a concrete example, when reasoning about the relationship between two variables, a prompt would give students general information about statistical dependence between variables, a hint would provide explanation of the exact phenomena under consideration (e.g., SC, AQ), and a bottom-out hint might ask for reasoning with a scatterplot (optimal graphical representation). Alternatively, reasoning with a 2-

Table 1: Examples of constructive reasoning coding applied for the analyses of posttest reasoning/code

Category	Sub-category	Sub-sub-category	Examples from data
Non-mathematical elaboration	Graphical	Complete	*Thinking for a good distribution that fit with this theory we can imagine a bar in the middle and nothing around. That means that all the people have the same degree of wealth. Looking at the plots we can already see that the distribution that seems what we have imagined is the normal distribution for the scenario A. We can also look at the standard deviation that confirm this reasoning*
		Not Complete	*Taking into account histogram with 50 bins, a better idea of distribution of wealth between citizens is given*
	Numerical	Complete	*By using a histplot, we see that for B there is no middle class, only rich and poor people =>not socialist.*
		Not Complete	*I add the values of each person and I divide by the number of person to find if the money is well distributed*
Mathematical elaboration	Graphical	Complete	*Datasets are almost identical specially in descriptive statistics but when we see plot of wealth distribution we can see that in B, there are more people with less wealth distribution specially after median and with similar reasoning we can say that as C is upper than B and A in most cases, it is the worst*
		Not Complete	*Linecharts show that C has the most wealth in the middle*
	Numerical	Complete	*Using the variance of each set, we can see that the values of dfA are much more centered around the mean (and thus a more egalitarian society). Followed by C then B*
		Not Complete	*comparing the median values of the different datasets*

D or 1-D histogram are examples of suboptimal RSMs. Here, information is lost because of binning and/or the lack of directly perceivable information about co-variation in the data.

Analytical Procedures

Due to dropout at various stages of the study (12%-57%), we applied standard multiple imputation (MI) procedures (n=5) to fill missing dataset values (Van Buuren, 2018). Discarding missing data may result in the complete cases being no longer representative of the target population, and consequently, estimates derived from them being subject to non-response bias. MI accounts for the process that created the missing data, and preserves uncertainty among relations in the data. Logistic regression and its variants (multinomial, ordered) were used for binary, nominal (>2 categories) and ordinal data respectively. Predictive mean matching was used to impute numeric data. Density plots of observed and imputed values were visually inspected for validity. Non-parametric statistics were used to see differences in ordinal posttest scores (e.g., Kruskal-Wallis tests, follow up Dunn tests). Multiple comparisons were adjusted using the Benjamini-Hochberg method. For non-significant results ($p > 0.05$), equivalence tests were performed to provide evidence for absence of a meaningful effect (Lakens, Scheel, & Isager, 2018). Here, the smallest effect size of interest was set within Cohen's d bounds of $\pm$ 0.2.

We also developed a coding scheme (Krippendorf's $\alpha >$ 0.7) for qualitative analyses of student's posttest reasoning and code, based on prior work (Chi, 2009; Kapur & Kinzer, 2009). First, we identified if reasoning was constructive (meaningful elaborations that went beyond what was presented). If yes, we identified if the elaborations were non-mathematical or mathematical. The former refers to elaborations that explain inferences leading up to the results, while the latter refers to elaborations that explicitly mention mathematical formalisms in words and/or in the code and base solution inferences on these formalisms. Next, for each kind of elaboration, we further checked if the elaborations comprised one or more graphical/numerical representation(s), meaning graphs, plots or other quantitative indices. Finally, we checked if these representation(s) were complete. Non-mathematical elaborations were coded as complete if all variables were set in relation to each other, and the result could be clearly derived from the elaboration. No information was missing and the connection between evidence and claim was fully explained using reasoning. Mathematical elaborations were coded as complete if all necessary methods in order to derive results were mentioned in words and/or presented in the code. Table 1 provides examples from the data.

Measures

Before the problem-solving phase, we collected student's prior knowledge using high school math scores as a proxy. No explicit pretest was conducted to prevent redundancy with the problem-solving phase. Based on prior literature on inter-individual factors that characterize heterogeneity in student's approach to FD and SD learning, we also included questionnaires assessing incoming profile variables like effort regulation (Pintrich et al., 1991), self-esteem (Jones, 1973), learning goal orientation (LGO) (Dweck, 1992) and attitude towards mistakes (ATM) (Leighton, Tang, & Guo, 2015). Effort regulation reflects a commitment to completing one's goals despite difficulties. High self-esteem triggers positive attributional style towards success and failure. An LGO disposition affects whether students view failures as learning opportunities. Finally, ATM, which includes the utility of making mistakes and induced affective reactions, enhances or impedes receptivity to failures. After the problem-solving phase, students answered task experience questionnaires, in line with PS-I preparatory mechanisms (Loibl et al., 2017).

These experiences included perceived awareness of knowledge gaps at the current moment (Glogger-Frey, Gaus, & Renkl, 2017), state curiosity about task actions and what they would like to know (Naylor, 1981), germane and extraneous cognitive load induced by problem-solving (Leppink, Paas, Van Gog, van Der Vleuten, & Van Merrienboer, 2014), and the experienced cognitive dissonance. Cognitive dissonance, defined as a state of discomfort associated with detection of conflicting concepts (Levin, Harriott, Paul, Zhang, & Adams, 2013), has not been studied in prior PS-I work because of lack of work on problematizing. Consistency of both incoming profile and task experience questionnaires was good for our dataset (McDonald's $\omega >$0.7). After the instruction phase, students solved an isomorphic and a non-isomorphic conceptual understanding posttest for each of the two task topics.

Results

Variable-centered Approach

We first performed variable-centered analyses to look at overall patterns of the impact of SD and FD preparatory activities on conceptual understanding in PF (figure 2).

Task topic SC For the SC topic, we found a significant omnibus effect of the experimental grouping on the

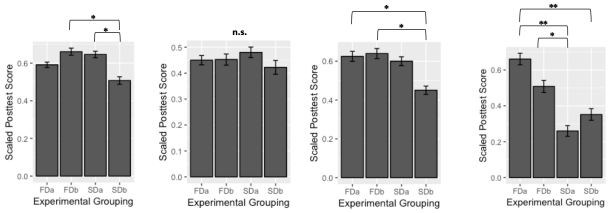

Figure 2: Scaled posttest scores with inferential error bars (L to R: SC non-isomorphic, SC isomorphic, AQ non-isomorphic, AQ isomorphic). Significant differences marked.

non-isomorphic conceptual understanding posttest ($\chi^2(3) = 11.73$, $p = 0.008$, Cohen's $d = 0.409$). The FD condition was better than the SD condition that offered the more-specific clue, in this case a hint describing the SC phenomena. However, the FD conditions were equivalent to the SD condition that offered the less-specific clue, in this case a prompt describing what statistical dependence among variables is. The two FD conditions here asked students to generate/reason with a correlation table and scatterplot matrix respectively, both of which reflect suboptimal numerical and graphical representations respectively. This is because they don't fully allow inferences on the nature of relationships (strength and/or meaningfulness) between dataset variables. No significant omnibus difference in scores on the isomorphic conceptual understanding posttest was observed across the four experimental conditions ($\chi^2(3) = 1.37$, $p = 0.712$, Cohen's $d = 0.174$). Equivalence testing suggested that the observed effect was neither statistically different from zero nor statistically equivalent, indicating insufficient data to draw conclusions.

Qualitative analysis suggested that for the non-isomorphic conceptual understanding posttest, the trend mirrored posttest scores. Students in the FD conditions had higher percentage of complete mathematical (32.1%, 38.7% >> 27.3%) and non-mathematical elaborations (44.7%, 56.7% >> 27.3%), compared to the SD condition with the more specific clue. Additionally, completeness of reasoning was almost identical between the FD condition and the SD condition with the less specific clue. However, for the isomorphic conceptual understanding posttest, student reasoning was often dominant in either complete mathematical or complete non-mathematical elaborations across the experimental conditions. The SD conditions had comparatively higher percentage of the former (38.9%, 45.8% >> 32.4%, 25%), while the FD conditions had comparatively higher percentage of the latter (33.3%, 48.8% >> 43.3%, 23.6%). This might be one reason why we saw no posttest score differences.

Task topic AQ For the AQ topic, we found significant omnibus effects of the experimental grouping on both the non-isomorphic ($\chi^2(3) = 10.84$, $p = 0.012$, Cohen's $d = 0.387$) and isomorphic ($\chi^2(3) = 20.16$, $p = 0.0001$, Cohen's $d = 0.586$) conceptual understanding posttest. Follow up pairwise comparisons suggested that scores for students in FD condition were greater than those in the SD condition with the more specific clue, in this case a bottom-out hint asking for scatterplot generation. However, the difference did not reach significance when comparing the FD condition and SD condition with the less specific clue, in this case a hint describing the AQ phenomena. The two FD conditions here asked students to generate/reason with a 2-D and 1-D histogram respectively. Both reflect suboptimal graphical representations.

We separated the coding of numerical and graphical representations to assess their independent usage in student reasoning. Qualitative analysis suggested that for the non-isomorphic conceptual understanding posttest involving graphical representations, students in the FD conditions had higher percentage of complete mathematical (72.2%, 68.7% >> 33.3%, 40%) and non-mathematical elaborations (44.4%, 37.5% >> 26.6%, 40%), compared to the SD conditions. This also held true for complete mathematical (27.7%, 37.5% >> 20%, 0%) and non-mathematical elaborations (27.7%, 31.2% >> 13.3%, 0%) involving numerical representations. For the isomorphic conceptual understanding posttest, a similar trend held for elaborations involving graphical representations. We did not see clear trends in qualitative differences in student reasoning for elaborations involving numerical representations, the less straightforward (and dominant) approach for this isomorphic question. Taken together, despite no posttest score differences between students who received FD scaffolds and the less-specific SD scaffold, there were salient differences in reasoning quality.

Person-centered Approach

We performed complementary person-centered analyses to go beyond an average FD or SD learning pattern (figure 3). The rationale here was to factor in the interactions among incoming student characteristics, in order to understand the impact of this heterogeneity on learning. We used latent profile analysis to first cluster students based on incoming profile variables like prior knowledge, effort regulation, learning goal orientation, self-esteem and attitude towards mistakes. This approach provides an elegant way to discover subgroups by "simultaneously" considering interactions among "more than one" incoming cognitive and motivational student characteristic. Non-parametric multivariate finite mixture models were used (Hickendorff, Edelsbrunner, McMullen, Schneider, & Trezise, 2017). A two-cluster solution (figure 4) reflected parsimonious fit to the data (based on model fit ($loglik = 530.41$), mixture distributions and visual inspection of mixture density plots when fitting more than two clusters).

Cluster assignments for students into these homogeneous subgroups were based on posterior probability distributions. These cluster assignments allowed us to then use this information for studying interaction effects (reported below). Statistically, we found one of these clusters (henceforth, Cluster_{high}) to have significantly higher scores on all of these incoming characteristics, compared to the other cluster (Cluster_{low}). Cluster_{low} reported higher extraneous cognitive load than Cluster_{high} ($W = 7319.5$, $p = 0.001$) after problem-solving. All other task experiences were statistically similar.

Task topic SC/AQ Not surprisingly, we did find that students in Cluster_{high} scored significantly higher than those in

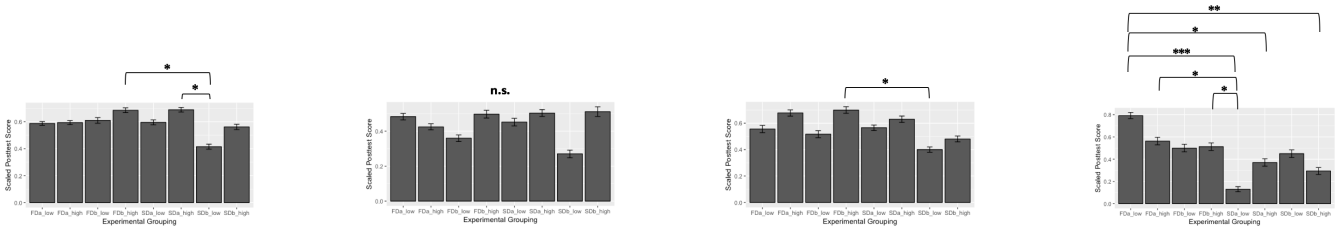

Figure 3: Scaled posttest scores with inferential error bars (L to R: SC non-isomorphic, SC isomorphic, AQ non-isomorphic, AQ isomorphic). Significant differences marked. *Low* and *High* represent students from Cluster$_{low/high}$ within a condition.

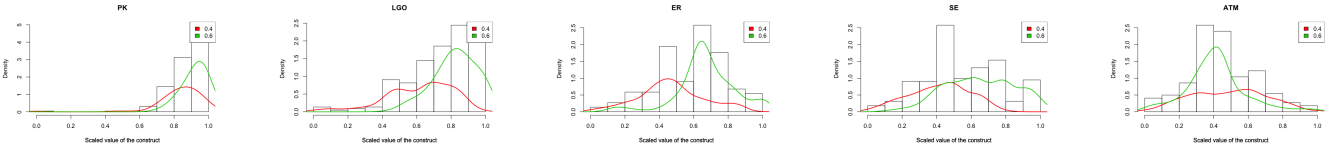

Figure 4: Mixture density distributions when clustering students based on incoming cognitive and motivational characteristics (L to R: Prior knowledge, Learning goal orientation, Effort regulation, Self-esteem, Attitude towards mistakes)

Cluster$_{low}$ on the non-isomorphic conceptual understanding posttest. This trend held for both SC ($W = 4885.5$, $p = 0.04$) and AQ task topics ($W = 4811.5$, $p = 0.02$). On the other hand, both Cluster$_{high}$ and Cluster$_{low}$ performed equally well on posttests scores for the isomorphic conceptual understanding question. Equivalence testing results were inconclusive.

Interaction We finally looked at impact of the interaction between experimental grouping and incoming student profile on posttest. Results suggested that the omnibus trend for difference in non-isomorphic conceptual understanding posttest still held for the SC topic ($\chi^2(7) = 17.16$, $p = 0.016$, Cohen's $d = 0.448$). Descriptively, students in the FD sub-groupings outperformed those in the SD sub-groupings. As before, the omnibus effect was still not significant for the isomorphic conceptual understanding posttest ($\chi^2(7) = 11.5$, $p = 0.118$, Cohen's $d = 0.294$). Equivalence testing showed 24/28 pairwise comparisons to be inconclusive.

For the more difficult topic AQ, again, as before, exposure to failure-driven scaffolds benefited students on both the non-isomorphic ($\chi^2(7) = 16.91$, $p = 0.017$, Cohen's $d = 0.442$) and isomorphic ($\chi^2(7) = 27.16$, $p = 0.0003$, Cohen's $d = 0.647$) conceptual understanding posttest. Descriptive trends for students in FD sub-groupings scoring higher than their counterparts in the SD sub-groupings still held. This also suggests that perhaps task difficulty and the extent to which student reasoning requires manipulation and integration of multiple representations, might be an important factor when looking at the relative efficacy of FD and SD scaffolds.

Underlying Mechanisms

We computed partial correlations between student's task experiences during the problem-solving phase and their posttest scores, controlling for experimental grouping (SD, FD) and incoming student profile (Cluster$_{high/low}$). For the more difficult topic (AQ), we saw positive associations of both isomorphic and non-isomorphic posttest scores with awareness of knowledge gaps ($\rho = 0.112^{+}$, 0.172*) and germane cognitive load ($\rho = 0.114^{+}$, 0.120^{+}). The correlation between these task experiences and posttest scores was not significant for the easier topic (SC). Experiencing higher state curiosity ($\rho = 0.158$*, 0.184**) and cognitive dissonance ($\rho = 0.193$**, 0.187**) was positively associated with only with non-isomorphic posttests, however for both SC and AQ topics. Finally, experiencing greater extraneous cognitive load was negatively associated with posttest scores for both SC (ρ = -0.243**, -0.236**) and AQ (ρ = -0.185**, n.s.) topics.

Manipulation Check and Design Implications

Students in every experimental condition had the opportunity to make two solution attempts (prior/post exposure to the scaffold) during the problem-solving phase. This design allowed us to assess the percentage of students who improved/degraded their solution across these two time points within the initial problem-solving. We computed a summary index S (ranging from -100 to 100) for each condition and task topic, by subtracting (i)ΔD, the percentage of students who degraded (got the right answer pre-scaffold, but wrong answer post-scaffold), from, (ii)ΔI, the percentage that improved (got the wrong answer pre-scaffold, but right answer post-scaffold). For the two FD conditions, we found S to be highly negative ($\Delta D > \Delta I$) for the more difficult task topic AQ (-72%, -47%), suggesting that the problematizing scaffold indeed pushed students towards explicit failure. For the easier task topic SC (-51%, -34%), S was still negative but comparatively lower in absolute terms.

Interestingly, for the two SD conditions, S was not positive or ΔI was $\not>$ than ΔD (as one might intuitively expect). Overall, despite S being lower in absolute terms compared to the FD conditions, it was still negative for both the AQ (-52%, -54%) and SC (-40%, -6%) task topics. This suggests that although explicit structuring prior to instruction led to greater net solution accuracy (compared to explicit problematizing), it was still not enough to push majority of student solutions to match the canonical answer. Taken together, these analyses show that students may not necessarily be prepared to receive explicit structuring during initial exploration, especially for difficult topics. It also opens up questions about re-calibrating the specificity of structuring scaffolds so that $\Delta I > \Delta D$.

Discussion and Conclusion

Table 2: posttest differences across experimental grouping

	Non-isomorphic conceptual understanding posttest	Isomorphic conceptual understanding posttest
Topic SC (Variable-centered)	$\chi^2(3) = 11.73$, $p = 0.008$ Cohen's $d = 0.409$	$\chi^2(3) = 1.37$, $p = 0.712$ Cohen's $d = 0.174$
Topic SC (Person-centered)	$\chi^2(7) = 17.16$, $p = 0.016$ Cohen's $d = 0.448$	$\chi^2(7) = 11.5$, $p = 0.118$ Cohen's $d = 0.294$
Topic AQ (Variable-centered)	$\chi^2(3) = 10.84$, $p = 0.012$ Cohen's $d = 0.387$	$\chi^2(3) = 20.16$, $p = 0.0001$ Cohen's $d = 0.586$
Topic AQ (Person-centered)	$\chi^2(7) = 16.91$, $p = 0.017$ Cohen's $d = 0.442$	$\chi^2(7) = 27.16$, $p = 0.0003$ Cohen's $d = 0.647$

To summarize, our results indicate the efficacy of FD over SD preparatory activities on student's conceptual understanding. We go beyond prior PS-I work by performing stringent comparisons between explicit ways of pushing students towards success and failure in problem-solving prior to instruction, and investigating their impact on learning. Overall, we found a significant main effect for experimental grouping on the non-isomorphic conceptual understanding posttest, with the FD conditions outperforming the SD condition with the more-specific clue, but not the SD condition with the less-specific clue. Posttest score similarity between the latter comparison indicates that FD and SD approaches might potentially offer two distinct but effective paths to learning. Nudging students to make them realize by themselves the extent to which their activated knowledge is (ir)relevant for solving the problem (we can have both SD and FD ways towards this end), is better than directing their activation of relevant prior knowledge (via a highly specific SD scaffold).

However, we also found that a comparatively higher percentage of students who received FD scaffolds demonstrated reasoning with complete mathematical or non-mathematical elaborations, indicating better quality of reasoning than students in the SD conditions. This result supports the idea that focusing on the pragmatic goal of performing the correct procedure (in presence of SD scaffolding) without appropriately articulating understanding (non-reflective work) can lead to fragile conceptual gains (Jonassen, 2010). We also found a significant main effect for the incoming student profile on the non-isomorphic conceptual understanding posttest, with students having high self-reported scores significantly performing better. There was no evidence for an interaction effect. Exposure to FD scaffolds had a greater impact on posttest scores for the more difficult topic (AQ). Finally, we found mechanistic task experiences to be positively associated with posttest scores (stronger associations for AQ task topic and for non-isomorphic posttests), controlling for experimental grouping and incoming profile.

What might explain the superiority of problematizing scaffolds over structuring scaffolds in the PS-I design? Although scaffolding for success might push for speed/accuracy to facilitate fluency in knowledge application for one form of independent performance (Schwartz, Sears, & Chang, 2007), both posttest scores and qualitative analysis of reasoning suggest that it does not guarantee improved conceptual understanding. Correct performance of a procedure scaffolded via structuring might stem from the lack of awareness and appreciation of long-term sub-optimality of a solution that works reasonably well in the short-term (Schwartz, Chase, & Bransford, 2012). The resulting quick/easy success may be insufficiently disruptive to challenge existing thought processes, and induce inattention when learning from instruction.

Existing meta-analysis of PS-I literature (Loibl et al., 2017) also suggests that students need to be made aware of the limitations to their knowledge (knowledge gaps). Further, we must instill in them a strong desire to know more about the canonical solution to fill these knowledge gaps. Finally, the learning design needs to facilitate understanding of which solutions don't work and why. In line with these vital pre-instructional goals, the suboptimal RSM generation strategy triggers "effortful activation" of prior knowledge conceptually relevant to the targeted learning concept.

By exposing students to additional exploration of the problem-space structure that doesn't necessarily lead to the canonical solution, suboptimal RSM generation provides support for meaningful variation in reasoning (Soderstrom & Bjork, 2015), which aids in improved conceptual understanding. Further, the uncertainty induced about consequences of partially-gained insights during solution revision is likely to trigger momentary curiosity driven by student's problem-solving experiences. One's own failed attempt is also likely to better prepare students for acquisition of negative knowledge regarding applicability conditions of solution strategies during instruction. Finally, at a methodological level, we see an improvement in effect sizes compared to a traditional variable-centered approach for both task topics (table 2). Complementary person-centered analyses provide a more accurate assessment of the impact of our PF intervention, since they factor in the differential benefits arising due to individual differences in SD and FD learning.

The scaffolding implemented in this work can be embedded into metacognitive tutors (Joyner & Goel, 2015) that deploy computer agents to imitate functional roles of teachers - "guides" to offer structuring, and "critiques" to problematize exploration. Limitations of this work stem primarily from the classroom time constraints. This was reflected, for e.g., in choice of datasets we used. For future work, we will design rich(er) datasets (that allow greater scope of inferences). The allocated time budget also led us to design one-step SD or FD scaffolds, and collect single task experience questionnaire after students finished solving problems on both topics (SC, AQ). Finally, optional university attendance resulted in considerable student dropout over the two study weeks, despite our efforts to mitigate this threat via participation reminders.

References

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).

Chase, C. C., & Klahr, D. (2017). Invention versus direct instruction: for some content, itsa tie. *Journal of Science Education and Technology*, *26*(6), 582–596.

Chi, M. T. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in cognitive science*, *1*(1), 73–105.

Chin, D. B., Chi, M., & Schwartz, D. L. (2016). A comparison of two methods of active learning in physics: inventing a general solution versus compare and contrast. *Instructional Science*, *44*(2), 177–195.

Dweck, C. S. (1992). Article commentary: The study of goals in psychology. *Psychological Science*, *3*(3), 165–167.

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.

Hickendorff, M., Edelsbrunner, P. A., McMullen, J., Schneider, M., & Trezise, K. (2017). Informative tools for characterizing individual differences in learning: Latent class, latent profile, and latent transition analysis. *Learning and Individual Differences*.

Holmes, N. G., Day, J., Park, A. H., Bonn, D., & 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).

Jonassen, D. H. (2010). *Learning to solve problems: A handbook for designing problem-solving learning environments*. Routledge.

Jones, S. C. (1973). Self-and interpersonal evaluations: esteem theories versus consistency theories. *Psychological bulletin*, *79*(3), 185.

Joyner, D. A., & Goel, A. K. (2015). Organizing metacognitive tutoring around functional roles of teachers. In *Cogsci.*

Kapur, M. (2011). A further study of productive failure in mathematical problem solving: Unpacking the design components. *Instructional Science*, *39*(4), 561–579.

Kapur, M. (2014). Comparing learning from productive failure and vicarious failure. *Journal of the Learning Sciences*, *23*(4), 651–677.

Kapur, M., & Bielaczyc, K. (2012). Designing for productive failure. *Journal of the Learning Sciences*, *21*(1), 45–83.

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

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 2515245918770963.

Leighton, J. P., Tang, W., & Guo, Q. (2015). *Developing and validating the attitudes towards mistakes inventory (atmi): A self-report measure.*

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.

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.

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).

Loibl, K., & Rummel, N. (2014). The impact of guidance during problem-solving prior to instruction on students inventions and learning outcomes. *Instructional Science*, *42*(3).

Mazziotti, C., Rummel, N., & Deiglmayr, A. (2016). Comparing students solutions when learning collaboratively or individually within productive failure. Singapore: International Society of the Learning Sciences.

Naylor, F. D. (1981). A state-trait curiosity inventory. *Australian Psychologist*, *16*(2), 172–183.

Pintrich, P. R., et al. (1991). A manual for the use of the motivated strategies for learning questionnaire (mslq).

Reiser, B. J. (2004). Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *The Journal of the Learning sciences*, *13*(3).

Roelle, J., & Berthold, K. (2016). Effects of comparing contrasting cases and inventing on learning from subsequent instructional explanations. *Instructional Science*, *44*(2).

Roll, I., Butler, D., Yee, N., Welsh, A., Perez, S., Briseno, A., ... Bonn, D. (2018). Understanding the impact of guiding inquiry: The relationship between directive support, student attributes, and transfer of knowledge, attitudes, and behaviours in inquiry learning. *Instructional Science*.

Schalk, L., Schumacher, R., Barth, A., & Stern, E. (2017). When problem-solving followed by instruction is superior to the traditional tell-and-practice sequence. *Journal of Educational Psychology*.

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

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

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.

Schwartz, D. L., Sears, D., & Chang, J. (2007). Reconsidering prior knowledge. *Thinking with data*, 319–344.

Sinha, T., & Kapur, M. (2019). When productive failure fails. In *Proceedings of the annual meeting of the cognitive science society.*

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

Van Buuren, S. (2018). *Flexible imputation of missing data*. Chapman and Hall/CRC.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*(4), 197–221.