When Productive Failure Fails

Tanmay Sinha, Manu Kapur
ETH Zürich, Switzerland

Abstract

Productive Failure (PF) is a learning design that intentionally designs for and uses failure in preparatory problem-solving for learning. Over the past decade, there has been growing evidence supporting the effectiveness of learning from PF. The purpose of this paper, however, is to critically examine evidence for when PF fails. We analyze 95 experimental comparisons from 57 studies reported in 44 articles into the extent to which they conform to PF design criteria. These criteria, as outlined in the original PF work, span the problem-solving activity, the participation structures, and the social surround. Results suggest lack of design fidelity as a critical factor for when PF fails to outperform alternative instructional approaches on conceptual knowledge and/or transfer.

Keywords: Direct Instruction; Productive Failure; Scaffolding

Introduction

The past decade has seen a growing body of evidence for the efficacy of Productive Failure (PF) for developing conceptual knowledge and transfer (for a review, see Kapur (2016); Loibl, Roll, and Rummel (2017)). PF comprises an initial problem-solving phase where learners generate and explore representations and solution methods (RSMs) to complex problems based on concepts they have not formally learnt yet, followed by an instruction phase where an expert or a teacher builds upon student-generated solutions to teach them the targeted concepts. According to PF, generating solutions to novel problems prior to instruction can help students learn better from the instruction, even if students fail to generate the correct solution in the problem-solving phase (Kapur, 2016). Thus conceived, PF can be seen as a subset of a general class of designs where problem-solving precedes instruction (or PS-I). It must be noted that not all PS-I designs are PF, but only those in which students generate multiple solutions but fail to generate the correct one.

In experimental comparisons, PF is typically compared with a design where students are initially given instruction on the targeted concepts, followed by problem-solving practice. Loibl et al. (2017) referred to this design as an Instruction-followed-by-Problem-Solving (I-PS) design. Findings in support of PF suggest that both PF and I-PS are similar in the development of procedural knowledge, but PF significantly outperforms I-PS in conceptual understanding and transfer (Kapur, 2016). Evidence comes not only from quasi-experimental studies conducted in the real ecologies of classrooms (e.g., Kapur (2012); Kapur and Toh (2013); Schwartz and Bransford (1998); Schwartz and Martin (2004)), but also from controlled experimental studies (e.g., M. S. DeCaro and Rittle-Johnson (2012); Kapur (2014); Loibl and Rummel (2014a); Roll, Aleven, and Koedinger (2011); Schmidt and Bjork (1992); Schwartz, Chase, Oppezzo, and Chin (2011)).

Although we now have substantial empirical evidence for when PF succeeds (Loibl et al., 2017), we argue it is equally important, if not more, to examine evidence when PF fails and delineate boundary conditions for how, when and why PF works. By success of PF, here we mean experimental comparisons in which PF significantly outperforms alternative instructional approaches (usually instruction followed by problem-solving (I-PS), but also scaffolded problem-solving followed by instruction (+PS-I), or a different preparatory activity followed by instruction (!PS-I)[1]). By failure of PF, here we mean experimental comparisons between PF and I-PS, PF and +PS-I, PF and !PS-I, where I-PS, +PS-I, !PS-I conditions significantly outperform PF on measures of either conceptual understanding or transfer.

At the same time, we also examine experimental comparisons with null results, that is, when there was no significant difference between PF and these three alternate experimental conditions. Although attribution of null effects to causal factors is not always straightforward, examining null effects may nevertheless shed light on the critical factors that confluence efficacy of PF. Bridging the gap between instructional decision-making and the science of learning from failure necessitates prescribing conditions under which positive or negative failure effects emerge and how to foster them.

Search Criteria

Our search process and the criteria for including and excluding comparisons for this analysis included articles in the Google Scholar databases that (i) cited either of the two seminal PF articles (Kapur, 2008; Kapur & Bielaczyc, 2012), and those that cited other key follow-up PF articles (Kapur, 2014, 2015, 2016), and (ii) reported experimental or quasi-experimental comparison between PF and I-PS, or between PF and +PS-I, or between PF and !PS-I; and (iii) assessed conceptual knowledge and/or transfer. Criteria i resulted in close to 700 articles as of 29th June 2018. Of these, 44 articles met criteria ii and iii. These 44 articles reported 57 studies and comprised 95 experimental comparisons[2]. Table 1 presents a breakdown of their demographic characteristics, with majority of the studies spanning Europe, North America and Asia, and covering mathematics concepts for 6th-10th graders. We also see evidence for PS-I work gradually expanding to different student populations at the post-graduate and professional levels within other STEM domains like physics, chemistry, biology, as well as within non-STEM domains like psychology and medicine.

Using a two-phase workflow, we now report key findings synthesized from these experimental comparisons. The first phase comprised a fidelity check for examining conformity of

[1] Exclamation (!) denotes [(NOT) Problem-solving], e.g., [Reading worked examples], [Problem posing], [Explanation generation]

[2] https://tinyurl.com/WhenPFfails

Table 1: Demographic characteristics of articles included in the review (Number of comparisons = 95)

	Variable of Interest	# of Comparisons (%)
1. Geographical distribution	Europe (Germany, Switzerland, UK)	30 (31.6%)
	North America (USA, Canada)	31 (32.6%)
	Asia (Singapore, Taiwan, India, Hong Kong, Saudi Arabia)	27 (28.4%)
	Australia	7 (7.4%)
2. Learner grade	6th - 10th graders	59 (62.1%)
	2nd - 5th graders	17 (17.9%)
	Undergraduates	16 (16.9%)
	Others (Postgraduates, Professionals)	3 (3.1%)
3. Targeted concept	Math (equivalence, geometry, fractions, variance, linear functions, central tendencies, least squares fitting, weighted averages, z-scores, statistics process control)	63 (66.3%)
	Physics (average speed, density, collision, electricity, mechanics)	16 (17%)
	Medical (dental hygiene, dental surgery)	4 (4.2%)
	Chemistry (solutions, atomic structure)	3 (3.1%)
	Psychology (memory)	2 (2.1%)
	Domain general skill (control of variables strategy)	2 (2.1%)
	Biology (genetics)	2 (2.1%)

PS-I implementations to PF design criteria (for detailed criteria definition, refer Kapur and Bielaczyc (2012)). A detailed breakdown of these PF fidelity check criteria for the current analyses is shown in table 2. Looking vertically across the table (from comparisons with positive results for PF to those with null and negative results for PF), the decrease in fidelity along many of the PF design criteria is striking. This suggests that our evidence base comprises a mixture of the original PF design as well as its low-fidelity versions. In the second phase, we explored additional reasons that could not be convincingly explained by fidelity check parameters alone. The rest of the article focuses on 44 of these 73 comparisons, 54.6% of which had significant negative ($p<0.05$) or null results ($p>0.05$) with I-PS as the comparison condition, 25% of which had negative or null results with +PS-I as the comparison condition, and remaining 20.4% of which had negative or null results with !PS-I as the comparison condition.

Negative Results for PF (compared to I-PS)

PF fidelity check revealed that most of the 7 experimental comparisons in this cluster (Loehr, Fyfe, & Rittle-Johnson, 2014; D. A. DeCaro, DeCaro, & Rittle-Johnson, 2015; Schalk, Schumacher, Barth, & Stern, 2017; Marei, Donkers, Al-Eraky, & van Merrienboer, 2017) considered affective draw of the problem (5/7), and provided evidence for multiple RSM generation during the initial problem-solving phase (5/7). However, what is striking is that in none of the comparisons did follow-up instruction build on failed or suboptimal learner generated solutions, or include group work as the participation structure. Since such consolidation and knowledge assembly is often a key component of PF (Kapur & Bielaczyc, 2012), we would not necessarily expect these low fidelity PF implementations to be better than I-PS comparison conditions. Other salient factors influencing results from these comparisons are described below.

First, learners with high performance orientation, who primarily seek to demonstrate ability, may view challenging task situations as a threat to this goal and withdraw their effort. Such learners are less likely than those with a learning-goal orientation disposition to focus on viewing failures as opportunities to learn, processing negative feedback as ways to improve performance, and experiencing positive emotions following failure (Dweck, 1992; Tulis & Ainley, 2011). Thus, there is no reason to believe that challenging exploratory problem-solving phase of PF might benefit them more so than an instruction-first approach (D. A. DeCaro et al., 2015).

Second, the presence of additional problem-solving practice following the PS-I routine allows learners to use the taught information immediately and integrate it with prior knowledge. Thus, PF can be expected to fail when the overall learning design lacks this practice activity, or, when the overall learning design includes this activity, but such an activity invokes application of procedural knowledge to solve problems and correct errors to a greater extent, rather than influencing processing and development of conceptual knowledge. Empirical evidence suggests that these negative effects were mitigated to some extent in a follow-up study (although not fully) when learners self-checked initial solutions immediately after instruction (Loehr et al., 2014).

Third, implementation-level details of preparatory problem-solving activities are important. PF can be expected to fail when the problem-solving phase comprises too loosely anchored instruction (e.g., an idealized contrasting case that represents a principle in an abstract and generic fashion, followed by self-explanation prompts). PF can, however also fail with relatively more anchored instruction (e.g., a grounded contrasting case that situates a principle in a specific context but also potentially contains (irr)relevant details, followed by self-explanation prompts).

In Schalk et al. (2017) for instance, idealized contrasting cases were operationalized by providing no labels for the axes of coordinate systems when introducing the concept of linear slopes in mathematics, while grounded cases had axes labeled with meaningful concepts (e.g., filling level in a rain barrel on the y-axis, and time in hours on the x-axis). Schalk et al. (2017) conjecture that although self-explanation prompts can help learners to abstract from the context provided in the grounded cases (Chi, De Leeuw, Chiu, & LaVancher, 1994), contextual details from the learning materials are likely to be preserved in the encoded knowledge representation (De Bock, Deprez, Van Dooren, Roelens, & Verschaffel, 2011). This can hamper transfer.

The detrimental effect of grounded cases might exist even if self-explanation prompts in the problem-solving phase are

Table 2: PF fidelity check criteria for the PS-I design, with 60 I-PS and 13 +PS-I and 22 !PS-I experimental comparisons. Results separated by positive, null and negative effects for PF. Table values show number (percentage) of comparisons. We describe an analyses of experimental comparisons with null and negative effects for PF in this paper.

Comparison condition	Effects for PF	Problems affording multiple RSMs	Evidence for multiple RSM generation	Affective draw of the problem	Group work as the participation structure	Building on learner solutions in Instruction
1. I-PS	Positive	36 (100%)	29 (80.5%)	32 (88.9%)	25 (69.4%)	23 (63.8%)
	Null	17 (100%)	8 (47%)	13 (76.4%)	9 (52.9%)	6 (35.3%)
	Negative	7 (100%)	5 (71.4%)	5 (71.4%)	0 (0%)	0 (0%)
2. +PS-I	Positive	2 (100%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)
	Null	7 (100%)	7 (100%)	7 (100%)	3 (42.8%)	1 (14.2%)
	Negative	4 (100%)	2 (50%)	4 (100%)	2 (50%)	2 (50%)
3. !PS-I	Positive	11 (100%)	6 (54.5%)	11 (100%)	5 (45.4%)	4 (36.3%)
	Null	5 (71.4%)	2 (28.5%)	5 (71.4%)	2 (28.5%)	2 (28.5%)
	Negative	4 (100%)	3 (75%)	4 (100%)	1 (25%)	2 (50%)

replaced by explicit invention prompts. From an instructivist point of view, the need to come up with unifying functional relations already makes the invention prompt inherently challenging. Addition of grounded cases can further overburden learners with unnecessary details. Experiencing increased extraneous load can negatively affect invention quality and subsequently transfer, placing learners in the PF condition at a disadvantage. More work is needed, however to understand relative efficacy of concrete or abstract preparatory activities.

Null Results for PF (compared to I-PS)

PF fidelity check revealed that most of the 17 experimental comparisons in this cluster (Schwartz & Martin, 2004; Belenky & Nokes-Malach, 2012; Matlen & Klahr, 2013; Loehr et al., 2014; Loibl & Rummel, 2014b; Fyfe, DeCaro, & Rittle-Johnson, 2014; D. A. DeCaro et al., 2015; Hsu, Kalyuga, & Sweller, 2015; Mazziotti, Loibl, & Rummel, 2015; Chase & Klahr, 2017; Tam, 2017; Marei et al., 2017; Newman & DeCaro, 2018) considered affective draw of the problem (13/17), about half of them provided evidence for multiple RSM generation during the initial problem-solving phase (8/17), while about one third of the comparisons included follow-up instruction building on learner generated solutions (6/17). This suggests moderate conformity to the PF design criteria, and calls for a nuanced understanding of the results.

While young learners (e.g., 2nd - 5th graders) may have insufficient prior knowledge about cognitive and metacognitive learning strategies to generate RSMs on their own (Mazziotti et al., 2015), adult learners with very high incoming mastery-approach orientation are likely to transfer regardless of the type of instruction. This is because the inventing activity in and of itself provides motivational impetus to learn the targeted concepts (Belenky & Nokes-Malach, 2012). These null results suggest that learners with such incoming cognitive or motivational profiles may not necessarily benefit from PF.

The nature of problem-solving task is an important factor as well. Tasks with high element interactivity (Sweller, 1988) have high expected error rate. As Loibl and Leuders (2018) suggest, revision of mental models following instruction for such tasks is contingent on whether or not learners spontaneously elaborate on erroneous solutions generated during initial problem-solving. As long as learners are prompted to explicitly compare and contrast their suboptimal solutions with the canonical solutions, they are likely to integrate negative knowledge in their repertoire of future problem-solving strategies. Consequently, there is no reason to suppose that such learners will benefit from problem-solving first (Hsu et al., 2015; Loibl & Leuders, 2018). While the sole impact of solution generation on the efficacy of PF is not yet clear, what is clearer is that the form of instruction matters (Loibl & Rummel, 2014b). Without instruction that compares and contrasts learner solutions with a canonical solution, PF can be expected to fail. Further, impact of the ordering of such instruction (before or after problem-solving) is less clear.

PF can also be expected to fail when the task provides no explicit feedback regarding what problem-solving actions are actually failures. Consequently, learners might not be in a position to use their awareness of knowledge gaps to consolidate information during the instruction phase (Matlen & Klahr, 2013). Finally, as Chase and Klahr (2017) suggest, when learning domain-general skills, the problem-solving phase in and of itself is less likely to provide implicit feedback about what goals to adopt during the inquiry process (that strongly impacts learning). For instance, learner's goals in pursuing inquiry might be scientific (finding out whether a variable impacts an outcome) or engineering-oriented (guarantying some desired outcome). In such situations, aligning learner's goals to a scientific one takes precedence over the relative ordering of the instruction phase in which this might happen.

Shifting focus to learner solutions, a key recurring factor for failure of PF is lack of evidence for learning to learn, i.e., spontaneous internalization of skills needed for application of domain-knowledge in novel situations. Gaining knowledge of how to perform a correct procedure after the consolidation phase of PF does not necessarily mean gaining high depth of understanding of the domain principle (Vollmeyer, Burns, & Holyoak, 1996; Schwartz, Chase, & Bransford, 2012; Soderstrom & Bjork, 2015). Self-regulated reasoning strategies (e.g., solution evaluation, unprompted self-explanation) require sufficient practice opportunities to get internalized (Schwartz & Martin, 2004; Tam, 2017). Finally, with respect to the overall learning design, PF is expected to fail or produce comparable effects to an I-PS design when the pretest targets concepts similar to the invention activity. Engaging learners in important exploratory learning processes such as prior knowledge activation, attention to knowledge gaps etc create redundancy with initial problem-solving phase of the PS-I setting, thus diluting effects (Newman & DeCaro, 2018).

Negative/Null Results for PF (compared to +PS-I)

PF fidelity check revealed that all the 11 experimental comparisons in this cluster (Kapur & Bielczyc, 2011; Holmes, Day, Park, Bonn, & Roll, 2014; Kim, Pathak, Jacobson, Zhang, & Gobert, 2015; Roelle & Berthold, 2016; Kuo & Wieman, 2016; Loibl & Leuders, 2018) considered affective draw of the problem, most of them provided evidence for multiple RSM generation during the initial problem-solving phase (9/11). However, about half of the comparisons used group work as the participation structure (5/11), and even fewer included follow-up instruction building on learner generated solutions (3/11). This suggests moderate conformity to the PF design criteria.

Evidence suggests that the extent to which activated prior knowledge is conceptually related to the targeted learning concept affects whether the failure resulting from it is productive. This can impact whether and when PF outperforms a scaffolded PS-I condition. If learners are scaffolded to detect high number of relevant similarities and differences in the contrasting cases during an initial problem-solving phase, this can lead them to focused elaboration/explanation regarding deep features of the problem after the instruction phase, resulting in improved conceptual understanding (Roelle & Berthold, 2016). Goal specificity research also suggests that the benefits of preparatory activities with low to medium goal specificity (as in the PS-I design) are contingent on affording opportunities for relevant prior knowledge activation, e.g., by guiding learners towards strategies that facilitate reasoning with the deep problem structure (Vollmeyer et al., 1996), or, by illustrating desirable sub-goals along a solution path that requires learners to focus on relevant task relationships (Miller, Lehman, & Koedinger, 1999). We describe such forms of scaffolded problem-solving in more detail below.

In the study by (Vollmeyer et al., 1996) for instance, explicit instruction in a systematic strategy (varying a single factor while holding other factors constant at zero) during the initial exploratory task fostered acquisition of the casual structure of a biological system. This was based on the premise that despite the presence of a nonspecific goal during the exploratory task, learners might not spontaneously make full use of effective rule-induction strategies. In the study by (Miller et al., 1999) where learners had to work in an exploratory micro-world to understand interactions of electrically charged particles, specializing the learning goal assisted learners in activating relevant prior knowledge. Illustrating a particular path and asking learners to arrange charged particles so that the moving charges would follow the illustrated path as closely as possible achieved this.

Richland and Simms (2015), more generally, have documented the importance of scaffolding exploratory problem-solving through a series of studies on induction within (non-) STEM domains. They emphasize explicit support in noticing the relevance of relational thinking, providing adequate processing resources to mentally hold and manipulate relations, and facilitating recognition of both similarities and differences when drawing analogies between systems of relationships. This is because learners may not spontaneously search for a common deep structure across problem instances.

Similar findings have been echoed in prior PS-I work (Schwartz et al., 2011; Kapur, 2015), which suggest that the benefits of prior knowledge activation such as noticing inconsistencies across multiple problem instances, encoding critical features from instruction etc are contingent on relevance of the activation. For instance, in an invention with contrasting cases study on the topic of density (Schwartz et al., 2011), students who recalled the deep structure of ratio from their invention activity were the ones who ultimately benefited from activating their prior knowledge on assessments of transfer. Scaffolding initial problem-solving as part of the PS-I design might then be one means to help learners activate relevant prior knowledge before receiving instruction.

Prior research on the mechanisms of errorful generation suggests that benefits are more likely when learners generate information semantically related to relevant task concepts and/or when subsequent feedback is related to these concepts (Clark, 2016). For e.g., in word-pair generation tasks, generations based on word stems or rhyming are unlikely to produce as much semantic activation, and do not show the beneficial effects of generation. Conceptual processing (guesses) afforded by error generation facilitate richer memory trace through ordered relations between errors and targets (leading to better recall and problem-solving performance), compared to, non-conceptual processing (lexical guesses) that is more likely to create retrieval noise without effortful semantic elaboration on part of the learner (Cyr & Anderson, 2015). Taken together, we can say that in absence of spontaneous task reasoning with relevant induction criteria (that can potentially be scaffolded within a +PS-I design), PF can fail.

However sometimes, even if task reasoning comprises relevant induction criteria, PF can be expected to fail if such task reasoning is then followed explicit instructions to come up with a unifying functional relation (how variables interact to produce a single quantitative result). Finding a very high number of similarities and differences in the contrasting cases (as part of initial task reasoning) can actually hurt posttest performance. Inventing can be expected to decrease learner's willingness to deeply process subsequent instruction because of clinging on to these self-generated suboptimal inventions (Johnson & Seifert, 1994), and valuing self-made products highly (Norton, Mochon, & Ariely, 2012). This can result in failure to recognize deficiency in problem-solving performance. Often, learner inventions fail to consider all factors necessary for developing the canonical solution, but focus only on subsets of these contrasting cases. In +PS-I work by Roelle and Berthold (2016), such detrimental effects increased as a function of the number of detected similarities and differences for which learners had generated inventions.

PF can fail if the delay caused in reaching an appropriate solution makes learners less interested and less self-efficient. As Glogger-Frey, Gaus, and Renkl (2017) found in their

work, this invoked feelings of knowledge insufficiency during preparation and consequently low confidence. With repeated failures, it becomes harder to perceive the value of engaging in good inquiry behaviors during the problem-solving phase because of lowered expectations and increased self-doubt (Ilgen & Hamstra, 1972), acceptance of absence of control (Mikulincer, 2013), susceptibility to demotivation and negative emotions like stress (LePine, LePine, & Jackson, 2004), and increased stability of future failure expectancies (Weiner, 1974). In +PS-I research conducted by Lee (2017) in physics, task failure in the form of circuit explosion (entire electrical circuit goes up in flames and a restart is required) was found to be negatively related to learning outcomes, perhaps because learners were not able to meaningfully grapple with the task complexity and lacked understanding of basic task elements. Prompts for metacognitive reflection did not help learners address these recurring failures.

Further, the temporal distance between the problem-solving and instruction phase matters. PF can be expected to fail if the instruction phase is temporally detached from all the conceptual exploration and reflection, compared to multiple smaller cycles of problem-solving and instruction happening closely together (Kim et al., 2015). The latter offers differentiated and redundant scaffolding opportunities (Tabak, 2004) to address the magnitude/diversity of knowledge assembly that learners need for understanding different conceptual task elements during the consolidation phase. Finally, PF can be expected to perform as well as +PS-I when cognitive support offered in the initial problem-solving phase is focused on principle-based guidance (covering definitions, conditions of applicability, relevant equations etc), as opposed to, being focused on clarifications and hints regarding correct solution steps, accuracy feedback etc. When learners have no or little relevant prior knowledge related to the target learning content, providing principle-based guidance during their initial problem-solving reduces extraneous cognitive load and in turn facilitates attention to critical task concepts.

Negative/Null Results for PF (compared to !PS-I)

PF fidelity check revealed that most of the 11 experimental comparisons in this cluster (Aleven, Koedinger, & Roll, 2009; Roll et al., 2011; Glogger-Frey, Fleischer, Grüny, Kappich, & Renkl, 2015; Kapur, 2015; Likourezos & Kalyuga, 2017; Newman & DeCaro, 2018) considered affective draw of the problem (9/11). However, about only half of these comparisons provided evidence for multiple RSM generation during the initial problem-solving phase (5/11). Further, only one third comparisons included follow-up instruction building on learner generated solutions (4/11) and used group work as the participation structure (3/11). This suggests low conformity to the PF design criteria. Comparison of such low fidelity versions of PF with !PS-I implementations indicates relatively lower extraneous load in !PS-I conditions as a key factor for the pattern of results. The !PS-I conditions usually include worked example followed by instruction, but sometimes also preparatory activities such as evaluating pre-designed solutions, problem-posing, reading/summarizing text etc followed by instruction.

One way to interpret the null results across these comparisons is by considering the relative contribution of different instructional activities and the socio-cognitive processes they trigger. As Kalyuga and Singh (2016) suggest, high(er) extraneous load for the PS-I condition is compensated by increase in intrinsic load (because of the diversity of instructional goals in the problem-solving phase such as prior knowledge activation, deep feature identification etc, as opposed to a solitary goal of solution schema acquisition). Also, PF learners experience motivational effects (acceptance of challenge, resolving conflict etc) that are different from those experienced by learners in a !PS-I condition (belief of success probability etc). Thus, one might conjecture the relative efficacy of PF over !PS-I implementations to depend on the balance between extraneous load and intrinsic load triggered by sequences of instructional tasks (that individually achieve different sub-goals). More research is needed along these lines.

Summary and Conclusion

We articulated factors representative of learner's situatedness relative to their problem-solving experiences to examine boundary conditions for failure of PF. PF (or more generally, PS-I) was compared with three alternate experimental conditions, (i) I-PS (instruction followed by problem-solving), (ii) +PS-I (scaffolded problem-solving followed by instruction), (iii) !PS-I (preparatory activity other than problem-solving followed by instruction). To summarize, our current analyses suggested low design fidelity (weak conformity to PF design criteria) as the starting point for when PF fails. However, deeper exploration into experimental comparisons with negative and null results for PF highlighted four important factors.

First, incoming cognitive and motivational characteristics (e.g., mastery orientation, self-regulation skills, inquiry goals) influence whether learners can be expected to benefit from PF. Second, nature of the problem-solving task (e.g., task difficulty/calibration to prior knowledge, triggered socio-cognitive processes, domain specificity, implicit task feedback) sheds further light into when PF can be expected to fail. Prior knowledge activation is a key cognitive mechanism that explains the beneficial effects of problem-solving based preparatory activities within the learning design of PS-I (Loibl et al., 2017). The boundary conditions explored in this work open up new research opportunities for developing variants of PF, or combining PF with other cognitively activating instructional methods (Hofer, Schumacher, Rubin, & Stern, 2018) for achieving stronger and more sustainable results. Such methods, which focus on learner's naïve concepts and beliefs as the starting point for knowledge construction and reorganization (Schneider & Stern, 2010) can include self-explanations, metacognitive questioning etc.

Third, learner solutions during the problem-solving phase (e.g., usage of relevant induction criteria, evidence for internalization, behavior rigidity) and the extent to which they

are scaffolded impacts learning from PF. Finally, nuances related to the overall PS-I learning design (e.g., redundancy of pretest, anchoring of initial problem-solving tasks, feedback in instruction phase, additional practice activities after instruction) matter for efficacy of PF activities over alternate designs. Although not exhaustive, these factors synthesized from studies around PF (and more broadly the PS-I literature) provide evidence-driven rationale for more careful design/labeling of future implementations. We hope this will spur lines of inquiry (e.g., see Sinha et al. (2019)) that design for balancing the incommensurable goals of learning versus performance (Soderstrom & Bjork, 2015), given the differential relationship of failure to these goals (Kapur, 2016).

References

Aleven, V., Koedinger, K., & Roll, I. (2009). Helping students know further–increasing the flexibility of students knowledge using symbolic invention tasks. In *Proceedings of the annual meeting of the cognitive science society.*

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

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., De Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive science*, *18*(3), 439–477.

Clark, C. M. (2016). *When and why does learning profit from the introduction of errors?* Unpublished doctoral dissertation, University of California, Los Angeles.

Cyr, A.-A., & Anderson, N. D. (2015). Mistakes as stepping stones: Effects of errors on episodic memory among younger and older adults. *Journal of experimental psychology: learning, memory, and cognition*, *41*(3), 841.

De Bock, D., Deprez, J., Van Dooren, W., Roelens, M., & Verschaffel, L. (2011). Abstract or concrete examples in learning mathematics? a replication and elaboration of kaminski, sloutsky, and heckler's study. *Journal for research in Mathematics Education*, *42*(2), 109–126.

DeCaro, D. A., DeCaro, M. S., & Rittle-Johnson, B. (2015). Achievement motivation and knowledge development during exploratory learning. *Learning and Individual Differences*, *37*, 13–26.

DeCaro, M. S., & Rittle-Johnson, B. (2012). Exploring mathematics problems prepares children to learn from instruction. *Journal of experimental child psychology*, *113*(4).

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

Fyfe, E. R., DeCaro, M. S., & Rittle-Johnson, B. (2014). An alternative time for telling: When conceptual instruction prior to problem solving improves mathematical knowledge. *British Journal of Educational Psychology*, *84*(3).

Glogger-Frey, I., Fleischer, C., Grüny, L., Kappich, J., & Renkl, A. (2015). Inventing a solution and studying a worked solution prepare differently for learning from direct instruction. *Learning and Instruction*, *39*, 72–87.

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.

Hofer, S. I., Schumacher, R., Rubin, H., & Stern, E. (2018). Enhancing physics learning with cognitively activating instruction: A quasi-experimental classroom intervention study. *Journal of Educational Psychology*.

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).

Hsu, C.-Y., Kalyuga, S., & Sweller, J. (2015). When should guidance be presented in physics instruction? *Archives of Scientific Psychology*, *3*(1), 37.

Ilgen, D. R., & Hamstra, B. W. (1972). Performance satisfaction as a function of the difference between expected and reported performance at five levels of reported performance. *Organizational Behavior and Human Performance*, *7*(3), 359–370.

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.

Kalyuga, S., & Singh, A.-M. (2016). Rethinking the boundaries of cognitive load theory in complex learning. *Educational Psychology Review*, *28*(4), 831–852.

Kapur, M. (2008). Productive failure. *Cognition and instruction*, *26*(3), 379–424.

Kapur, M. (2012). Productive failure in learning the concept of variance. *Instructional Science*, *40*(4), 651–672.

Kapur, M. (2014). Productive failure in learning math. *Cognitive Science*, *38*(5), 1008–1022.

Kapur, M. (2015). The preparatory effects of problem solving versus problem posing on learning from instruction. *Learning and instruction*, *39*, 23–31.

Kapur, M. (2016). Examining productive failure, productive success, unproductive failure, and unproductive success in learning. *Educational Psychologist*, *51*(2), 289–299.

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

Kapur, M., & Bielczyc, K. (2011). Classroom-based experiments in productive failure. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 33).

Kapur, M., & Toh, P. L. L. (2013). Productive failure: From an experimental effect to a learning design. *Educational design research–Part B: Illustrative cases*, 341–355.

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

Kuo, E., & Wieman, C. E. (2016). Toward instructional design principles: Inducing faradays law with contrasting

cases. *Physical Review Physics Education Research*, *12*(1), 010128.

Lee, A. (2017). *Productive responses to failure for future learning*. Columbia University.

LePine, J. A., LePine, M. A., & Jackson, C. L. (2004). Challenge and hindrance stress: relationships with exhaustion, motivation to learn, and learning performance. *Journal of Applied Psychology*, *89*(5), 883.

Likourezos, V., & Kalyuga, S. (2017). Instruction-first and problem-solving-first approaches: alternative pathways to learning complex tasks. *Instructional Science*, *45*(2).

Loehr, A. M., Fyfe, E. R., & Rittle-Johnson, B. (2014). Wait for it... delaying instruction improves mathematics problem solving: A classroom study. *The Journal of Problem Solving*, *7*(1), 5.

Loibl, K., & Leuders, T. (2018). Errors during exploration and consolidation - the effectiveness of productive failure as sequentially guided discovery learning. *Journal für Mathematik-Didaktik*, *39*(1), 69–96.

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. (2014a). The impact of guidance during problem-solving prior to instruction on students inventions and learning outcomes. *Instructional Science*, *42*(3), 305–326.

Loibl, K., & Rummel, N. (2014b). Knowing what you don't know makes failure productive. *Learning and Instruction*, *34*, 74–85.

Marei, H. F., Donkers, J., Al-Eraky, M. M., & van Merrienboer, J. J. (2017). The effectiveness of sequencing virtual patients with lectures in a deductive or inductive learning approach. *Medical teacher*, *39*(12), 1268–1274.

Matlen, B. J., & Klahr, D. (2013). Sequential effects of high and low instructional guidance on childrens acquisition of experimentation skills: Is it all in the timing? *Instructional Science*, *41*(3), 621–634.

Mazziotti, C., Loibl, K., & Rummel, N. (2015). Collaborative or individual learning within productive failure: Does the social form of learning make a difference? International Society of the Learning Sciences, Inc.[ISLS].

Mikulincer, M. (2013). *Human learned helplessness: A coping perspective*. Springer Science & Business Media.

Miller, C. S., Lehman, J. F., & Koedinger, K. R. (1999). Goals and learning in microworlds. *Cognitive Science*, *23*(3).

Newman, P., & DeCaro, M. (2018). How much support is optimal during exploratory learning? In *Proceedings of the 40th annual conference of the cognitive science society.*

Norton, M. I., Mochon, D., & Ariely, D. (2012). The ikea effect: When labor leads to love. *Journal of consumer psychology*, *22*(3), 453–460.

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

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

Roll, I., Aleven, V., & Koedinger, K. (2011). Outcomes and mechanisms of transfer in invention activities. In *Proceedings of the annual meeting of the cognitive science society.*

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*.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological science*, *3*(4), 207–218.

Schneider, M., & Stern, E. (2010). The cognitive perspective on learning: Ten cornerstone findings. *The nature of learning: Using research to inspire practice*, 69–90.

Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. *Cognition and instruction*, *16*(4), 475–5223.

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.

Sinha, T., Kapur, M., West, R., Catasta, M., Hauswirth, M., & Trninic, D. (2019). Impact of explicit failure and success-driven preparatory activities on learning. 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.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive science*, *12*(2), 257–285.

Tabak, I. (2004). Synergy: A complement to emerging patterns of distributed scaffolding. *The journal of the Learning Sciences*, *13*(3), 305–335.

Tam, K. (2017). Examining productive failure instruction in dental ethics.

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.

Vollmeyer, R., Burns, B. D., & Holyoak, K. J. (1996). The impact of goal specificity on strategy use and the acquisition of problem structure. *Cognitive Science*, *20*(1).

Weiner, B. (1974). *Achievement motivation and attribution theory*. General Learning Press.