Productive Failure and Subgoal Scaffolding in Novel Domains

Dar-Wei Chen[1](✉) and Richard Catrambone[2]

[1] Soar Technology Inc., 12124 High Tech Ave,
Suite 350, Orlando, FL 32817, USA
darwei.chen@soartech.com
[2] Georgia Institute of Technology, Atlanta, GA 30332, USA

Abstract. The assistance dilemma asks how learning environments should "balance information or assistance giving and withholding" (Koedinger and Aleven 2007, p. 239). Minimal guidance (MG) methods posit that students learn best when exploring problems freely, while direct instruction (DI) methods provide canonical solutions early on to streamline students' efforts (problems later). Each method type provides unique benefits, but both are important (Schwartz and Martin 2004) and not easily delivered together. A relatively new MG-based method called "productive failure" (PF) is hypothesized to capture both sets of benefits by requiring students to struggle through problems early on and revealing canonical solutions afterward (Kapur 2008). Students using PF are hypothesized to more effectively transfer and retain information because balancing heuristics and formal knowledge produces diverse solution attempts (diSessa and Sherin 2000) and struggling during exploration pushes students to fill knowledge gaps (Kulhavy and Stock 1989). In the present studies, participants learned to perform tasks in two domains, cryptarithmetic (more traditional) and Rubik's Cube (psychomotor, less traditional) while using either PF or DI. Analyses revealed that (A) PF participants did not outperform DI participants on either immediate post-tests or retention tests, although they did report being more exploration-oriented and trying more unique strategies, (B) subgoal labels increased learning, but only for the relatively novel Rubik's Cube domain (and they sometimes increased workload in cryptarithmetic, surprisingly), and (C) effects of subgoal labels did not change with instruction type. Future research should determine how PF methods can be scaffolded to foster exploration mindsets and diverse solutions.

Keywords: Productive failure · Educational methods · Scaffolding · Assessment · Training transfer · Retention · Subgoals

1 Introduction

For many years, education researchers have debated a seemingly simple question called "the assistance dilemma," which can be summarized as: "How should learning environments balance information or assistance giving and withholding to achieve optimal student learning?" (Koedinger and Aleven 2007, p. 239). The answer to this question has the potential to shape future instructional design in fundamental ways, but no

R. A. Sottilare and J. Schwarz (Eds.): HCII 2019, LNCS 11597, pp. 282–300, 2019.
https://doi.org/10.1007/978-3-030-22341-0_23

consensus has been reached thus far. For now, two categories of instructional methods dominate the debate. Traditional methods that provide canonical instruction early on and utilize problem-solving as application practice are called "direct instruction" (DI), while "minimal guidance" (MG) methods require learners to discover information through guided exploration and problem-solving, instead of receiving canonical instruction.

Although MG and DI methods are pedagogically different, they are similar in that they both strive to help students avoid struggle and failure (i.e., being unsuccessful in producing canonical solutions) while learning; both types of methods provide various levels of scaffolding to reduce learner struggle and failure, ostensibly because struggle and failure ultimately do more harm than good. However, a relatively new method called "productive failure" (PF; e.g., Kapur 2008) is hypothesized to leverage struggle and failure for unique learning benefits. In PF, learners attempt problems first before receiving canonical instruction and it is hypothesized that as a result, they will potentially be abler to (A) solve transfer problems, (B) retain knowledge past immediate comprehension tests, (C) know *why* a given solution is correct, as opposed to just knowing *that* it is correct, and (D) identify their own gaps in knowledge, among other benefits. Furthermore, given that PF is an exploration-based method with canonical instruction implemented, learners using PF are hypothesized to reap benefits usually associated with either minimal guidance (e.g., self-generated concepts) and direct instruction (e.g., streamlining of attention and resource allocation). The experiments described here tested the "productive failure" hypothesis and aim to provide new perspective to existing learner assistance approaches as well.

1.1 "Minimal Guidance" Model

Productive failure methods are based, in part, on a variety of existing minimal guidance methods, but PF is hypothesized to improve on each of those methods in some fashion.

- **Discovery learning.** An early instantiation of minimal guidance was "discovery learning," in which students freely explore domains and material for themselves to create governing insights about the world (Anthony 1973), often without concrete goals in mind.
- **Constructivism.** Learners in constructivist environments are hypothesized to build "conceptually functional representations of the external world" that are not necessarily unique to themselves (Jonassen 1991, p. 61). Therefore, while the basic pedagogical premise of constructivism is similar to that of discovery learning (i.e., active construction of meaning), a conceptual difference is that in discovery learning, students are hypothesized to instead construct their own unique representations of the world.
- **Impasse-driven learning.** Impasse-driven learning is one of the first methods to implement struggle and failure to a large extent; impasses are defined by VanLehn et al. (2003, p. 220) as situations in which a student is stuck, "detects an error, or does an action correctly but expresses uncertainty about it." The governing principle of impasse-driven learning is that impasses are effective in helping learners adopt learning-oriented mindsets, which cause them to be more likely to search their

memories, examine the environment, or ask nearby people, etc. in attempts to discover what they do not understand (VanLehn et al. 2003). After students reach impasses, tutors are to provide explanations soon after when students are not able to.

No matter the specific instantiation, MG methods are hypothesized to mitigate working memory constraints by encouraging learners to connect new information with prior long-term knowledge (Kapur and Bielaczyc 2011) during unstructured problem-solving periods. These connections increase the chances that new information is understood at a deeper level than if it was learned via DI, where the new information is often stored in working memory and available in external memory.

1.2 "Direct Instruction" Model

Opposite minimal guidance in the learner assistance debate are direct instruction methods, which generally guide students strongly and limit exploration. The worked example is considered "the epitome of strongly guided instruction" and "provides some of the strongest evidence for the superiority of directly guided instruction over minimal guidance" (Kirschner et al. 2006, p. 80). Worked examples are hypothesized to streamline attention to the most important parts of problems, reducing problem-solving search and thus lower working memory loads (Kirschner et al. 2006). For most learners, and novices in particular, this streamlining is key because they do not possess the relevant schemas with which to integrate new information and prior knowledge, and therefore cannot construct new schemas that are durable (Rourke and Sweller 2009). When unguided, many novices often resort to methods such as trial-and-error which are burdensome on working memory, causing it to be unavailable for contributing to long-term memory (Kirschner et al. 2006). If working memory is occupied with tasks such as trial-and-error or problem-solving search, unguided students will not be able to use working memory to learn, and they could therefore potentially search problem spaces for long periods without adding to long-term memory (Sweller et al. 1982). Learners can also sometimes lean too much on pre-existing knowledge to explore a domain (as opposed to devising learning goals), which can then lead to flawed conclusions (Wineburg and Fournier 1994). Direct instruction can be instantiated in many ways: Lectures, models, videos, presentations, demonstrations, as well as the aforementioned worked examples (Clark et al. 2012).

1.3 Solving the Assistance Dilemma Through "Productive Failure"

A growing body of literature posits that the productive failure methodology can help students learn in ways that achieve the objectives of both minimal guidance and direct instruction (e.g., Kapur 2011); that is, the "MG vs. DI" debate might be a false choice.

On a high level, productive failure requires students to invent solutions to presented problems first (in the "generation period") before receiving canonical instruction ("consolidation period"), thereby reversing the traditional order of these two teaching elements in DI. This order leads to struggle (and ultimately, failure) early on in the learning process, but there often exists "a latent productivity in what initially seemed to be failure" (Kapur 2008, p. 379). The generation effect, "which refers to the long-term benefit of generating an answer, solution, or procedure versus being presented that answer, solution, or procedure" (Bjork and Bjork 2011), could explain this latent productivity, in part. The ensuing canonical instruction then serves to combat the "negative transfer" (Bransford and Schwartz 1999) that often plagues minimal-guidance methods. It should be noted, however, that PF students do receive some basic domain information before entering the generation period, which lessens the probability of unproductive failures in which students attempt solutions that are too irrelevant to yield any valuable information.

Most MG methods employ scaffolding so that learners can avoid failure, ostensibly because it will hinder learning; however, failure is embraced and explicitly designed into the PF process through the use of problem-solving early on (generation period), and difficult ill-structured problems in particular are frequently used. In practice, scaffolding is withheld and "solution features" are deliberately made inconspicuous in PF so learners will be unlikely to guess canonical solutions, instead being encouraged to lean on heuristics and prior knowledge to generate solutions (Loibl and Rummel 2014a).

After initial problem-solving, canonical instruction follows for learners to fill in the rest of their understanding and remedy any mistakes they made. Sometimes, an initial assessment is implemented first immediately after the initial problems to ensure more concrete failure. Each of the following sections summarizes a key component of the productive failure hypothesis.

Heuristics Plus Formal Knowledge. In minimal-guidance environments, learners are led to utilize prior knowledge and heuristics during problem-solving, thereby mitigating some working memory constraints (Kapur and Bielaczyc 2011) on the whole, even if searching problem spaces also increases learners' working memory burdens somewhat (Sweller 1988). In the event that some learners do encounter higher cognitive demands in PF, they also often report feeling more engaged because of the autonomy they are afforded during initial problem-solving (diSessa et al. 1991). This prior knowledge activation is crucial for helping learners connect new material with long-term knowledge, which enables better encoding and assembling of schemas (Hiebert and Grouws 2007) as well as better transferability and durability of learning (Kapur 2008).

The blending of heuristics, prior knowledge, and formalized canonical instruction allows PF methods to provide benefits that MG and DI alone cannot. For example, PF students are more likely to generate relatively large amounts of diverse solutions for novel problems (diSessa and Sherin 2000), a hallmark of how experts attempt problems (Clement 1991). Through these diverse solution attempts, students are expected to develop the ability to extrapolate new information to other contexts (procedural flexibility; Gorman et al. 2010). Another hypothesized benefit is the priming of students to solve transfer problems later using the relative wealth of available information (prior knowledge, heuristics, canonical instruction), even if the information is not germane to any given initial problem (Bransford and Schwartz 1999).

A fair question regarding the above information might be whether DI methods can also achieve results similar to PF, given that many of them also implement canonical instruction and problem-solving. The key difference is that in productive failure, students use problem-solving to "assemble or structure key ideas and concepts while attempting to represent and solve the ill-structured problems" (Kapur et al. 2010, p. 1722). However, in direct instruction, problems are used "not as vehicles for making discoveries, but as a means of *practicing* recently-learned content and skills" (Clark et al. 2012, p. 6). As a result, students in DI are less likely to blend heuristics and formal knowledge, and more likely to receive formal knowledge and merely re-activate it when solving problems, leading to transferability that is not as robust. In contrast, PF students are led to use heuristics and prior knowledge during initial problem-solving (before receiving canonical instruction to remedy gaps in understanding), which ensures that both knowledge types are activated while learning. The order of material presentation is the key difference.

Failure-Related Cognition. "Expectation failure" is the idea that learning is most successful when the outcome expected by a student from the domain does not, in fact, occur (Schank 1997). Key principles of expectation failure include:

- Learners are less likely to develop creative solution attempts if environment is too controlled and failures are therefore not possible
- Learners are predisposed to explaining occurrences in the domain and adjusting their mental models to avoid being surprised by similar events
- For expectation failures to be most effective, they must occur during initial/practice problem-solving (more likely to be activated in future problems)

The key function of expectation failures is exposing learners to gaps in their understanding and eliciting learners' natural misunderstanding-induced curiosity in the material. In these situations, learners are more driven to fill knowledge gaps on their own (e.g., studying feedback), particularly when discrepancies between solution attempts and canonical solutions are wide (Kulhavy and Stock 1989). Due to the "problem-solving prior" instructional order, PF methods are particularly conducive to learners producing initial solution attempts that are discrepant from canonical solutions. Expectation failures also disrupt learners' stability bias, the overconfident belief that currently-accessible information will remain just as accessible in the future (Kornell and Bjork 2009). Chi's (2000) theory of the imperfect mental model also accords with the notion that failure can be effective and essential for learning; in short, the theory states that learning is done through updates to one's own mental models and that self-explaining, in particular, is an efficient way for learners to update their own models according to their own needs.

Furthermore, when experiencing failures and ensuing canonical instruction, learners will also tend to identify reasons that a solution is plausible and why non-canonical solutions do not always work, which improves their capacity for transfer to novel situations (Kapur and Lee 2009). Comparing invented solutions and canonical solutions aids in the encoding of critical conceptual features and selecting relevant problem-solving procedures, even when performing transfer tasks (Siegler 2002). For example, when students were allowed to observe the consequences of entering incorrect spreadsheet formulas, as opposed to being corrected immediately upon entering an incorrect formula, they achieved higher scores on transfer tasks than immediately-corrected students (Mathan and Koedinger 2003).

Immediate Performance vs. Enduring Learning. "Desirable difficulties" (Bjork and Bjork 2011), even if not severe enough to consistently induce failure, can still induce decreased immediate performance and PF-related learning benefits in the long term. Examples of these difficulties include environmental factors (e.g., interface clutter; Fiore et al. 2006), training variation (e.g., practicing tasks that are adjacent to the target task; Kerr and Booth 1978), practice scheduling (e.g., interleaved schedule produces better retention than blocked schedule; Shea and Morgan 1979), and secondary tasks (adding relevant concurrent secondary task improves test performance; Young et al. 2011).

The goal of any instructional method should be learning, which can be defined as "permanent changes in comprehension, understanding, and skills of the types that will support long-term retention and transfer" (Soderstrom and Bjork 2015, p. 176). Learning is a separate observed variable from immediate performance, which is a possibly temporary measure that can be an unreliable indicator of learning (Soderstrom and Bjork 2015). Many instructional methods focus on producing immediate performance improvements, but some evidence indicates that immediate performance is not indicative of long-term retention and/or transfer, which is perhaps more important (e.g., Schmidt and Bjork 1992).

When learners demonstrate strong immediate performance, they could be merely exhibiting retrieval strength, which is recall activated in particular contexts; however, durable learning is a function of storage strength, which comprises the depths to which the material is associated with prior knowledge (Bjork and Bjork 2011). Increasing storage strength is most efficiently done through information retrieval (as opposed to information review) because the creation of "new routes" to information inherently activates previous knowledge as well (Carrier and Pashler 1992). The observation that enduring learning and immediate outward performance improvement can be uncorrelated is seen in research ranging from maze rats (rats' abilities to finish mazes improve after ostensibly random wandering; Blodgett 1929) to statistics classes (students who invented solutions and received canonical instruction later outperformed DI students; Schwartz and Martin 2004). Furthermore, methods that aim to improve immediate performance can actually undermine enduring learning: For example, frequent and/or specific feedback, a common DI component, often helps students complete test problems that are similar to the ones they practiced, especially if tested soon after instruction. However, learners that receive the crutch of immediate and frequent feedback are shielded from creating generalizable problem-solving strategies, an important skill that is developed in those that are forced to struggle without immediate feedback (Cope and Simmons 1994).

1.4 Examining Subgoal Scaffolding in Productive Failure

Many of the PF studies to this point have required learners to complete initial problem-solving (the "generation period") without scaffolding of any sort, perhaps because this arrangement increases the chances of failure and the learner reaping the benefits associated with failure. When no scaffolding structure is present, one potential concern is that learners might not fail in constructive ways, which could then lead to difficulty during canonical instruction because learners will have strayed “off course” to varying extents. Therefore, it is possible that PF methods could be even more optimal for learning with the implementation of some scaffolding, especially those scaffolding mechanisms that provide just enough guidance to ensure that failures are indeed productive (i.e., help students unearth fundamental truths about the domain).

A few PF studies have implemented scaffolding during the generation period, but there are many more scaffolding mechanisms to be examined with regards to interactions with PF, some of which might produce better learning than non-scaffolded PF methods. The scaffolding mechanism chosen for manipulation in the current study is “subgoals,” which are labels for functional groupings of steps that can help learners recognize fundamental components of a problem (Catrambone 1998). Subgoals are a promising scaffolding mechanism for PF because they can potentially alleviate one of the major weaknesses in PF methods, which is the possibility that learners might fail unproductively by misunderstanding the deep structure of a given problem space.

1.5 General Overview of Current Study and Hypotheses

The experiments in the present study compared the effectiveness of productive failure and direct instruction in two domains that have not been examined before in this PF context. In Experiment 1, participants learned about cryptarithmetic, a domain that functions like the traditional academic domain of algebra and is somewhat similar to physics and math domains that have been used in past PF studies, but is more likely to be unfamiliar to participants (example problem: OOOH + FOOD = FIGHT). The tasks inherent in this domain (deducing variable values, logical reasoning, etc.) allow for reasonable comparison of the results to those from existing PF studies, which have centered mostly on STEM domains. In Experiment 2 (which was procedurally identical to Experiment 1), participants learned about solving the first layer of the Rubik’s Cube, a spatially-oriented task that requires some psychomotor coordination. The generalizability of PF methods to non-traditional domains were tested in this experiment. Experiment 2 provided an opportunity to examine whether Experiment 1 findings replicated or whether the effects of the manipulations might depend on how academic in nature the domain is. The specific methodological details used in these experiments can be found in the next section.

2 Method – Experiment 1 (Cryptarithmetic)

2.1 Participants

A meta-analysis of productive failure studies (Chen 2016) found that PF methods have produced, on average, a performance improvement of about 0.66 SD in deep conceptual knowledge when compared to direct instruction methods, and because PF was hypothesized to improve this kind of generalizable knowledge (as opposed to performance on procedurally-similar tasks), this effect size drove the power analysis used to determine the sample size in this study. To achieve 80% power and 5% Type I error rate when searching for an effect of this size, 64 participants were used. These participants were recruited through the online SONA research participation system at the Georgia Institute of Technology and compensated with class credits for their time. All students at the Institute qualified for the experiment except for those who had prior experience in systematically solving cryptarithmetic problems.

2.2 Experimental Design

Experiment 1 was a laboratory experiment in which all participants were required to learn how to solve basic cryptarithmetic addition problems involving two numbers. The two manipulated independent variables that will be covered in this paper are:

- Instruction type (between subjects): productive failure or direct instruction
- Subgoal labels (between subjects): subgoal labels were provided or withheld

All variables were fully crossed to form a factorial design for the experiment. Observed dependent measures included immediate task performance (near transfer, medium transfer, far transfer), retention task performance after a one-week break (near transfer, medium transfer, far transfer), and several secondary assessments that could predict task performance (e.g., workload, number of solution methods generated).

2.3 Procedures

Table 1 summarizes, in order, the procedures that participants completed during the experiment and some associated details.

3 Method – Experiment 2 (Rubik's Cube)

Procedures for Experiment 2 were identical to those in Experiment 1 except for the domain (see Table 2); participants in Experiment 2 learned how to solve the first layer of the Rubik's Cube.

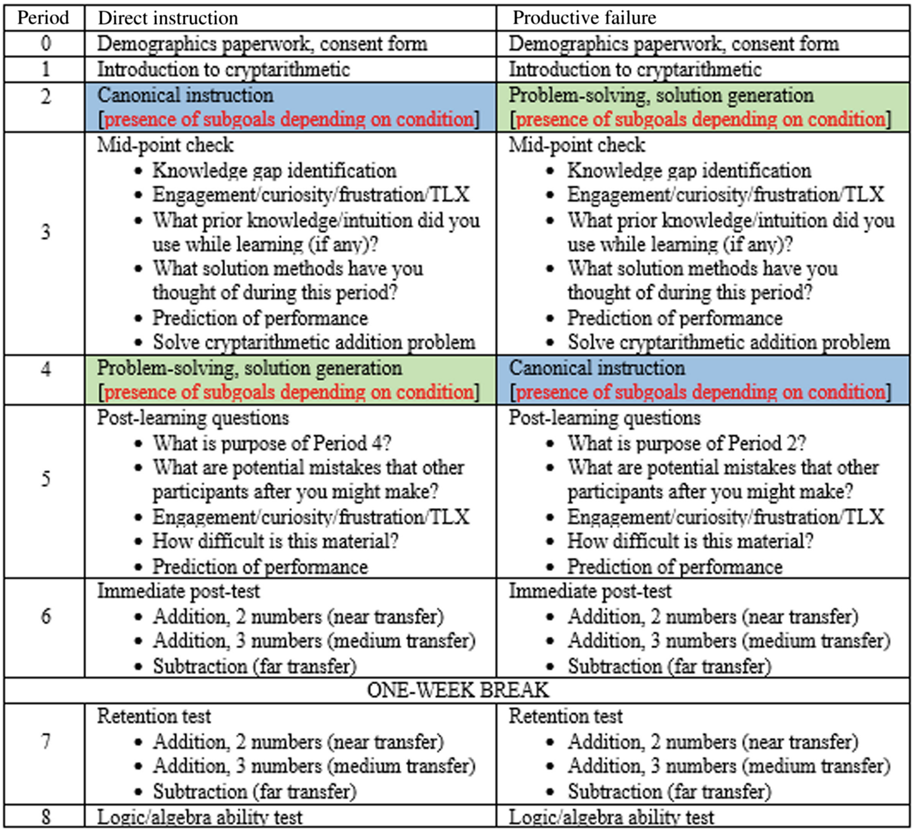
4 Results and Discussion

4.1 Instruction Type Main Effects (Immediate Post-test)

A general linear model (GLM) was created to analyze how the manipulated independent variables affected immediate post-test scores in both domains. For each individual problem type as well as overall test score, the data indicated that there was generally no significant difference between productive failure and direct instruction, except for one instance that is likely a random outlier given the pattern of the other results. Table 3 outlines these results (maximum possible test score is 100%).

In the realm of near-transfer test problems, it was not expected that productive failure would produce significantly better task performance than direct instruction, especially when the problems were administered immediately after learning has

Table 1. Summary of Experiment 1 procedures (cryptarithmetic)

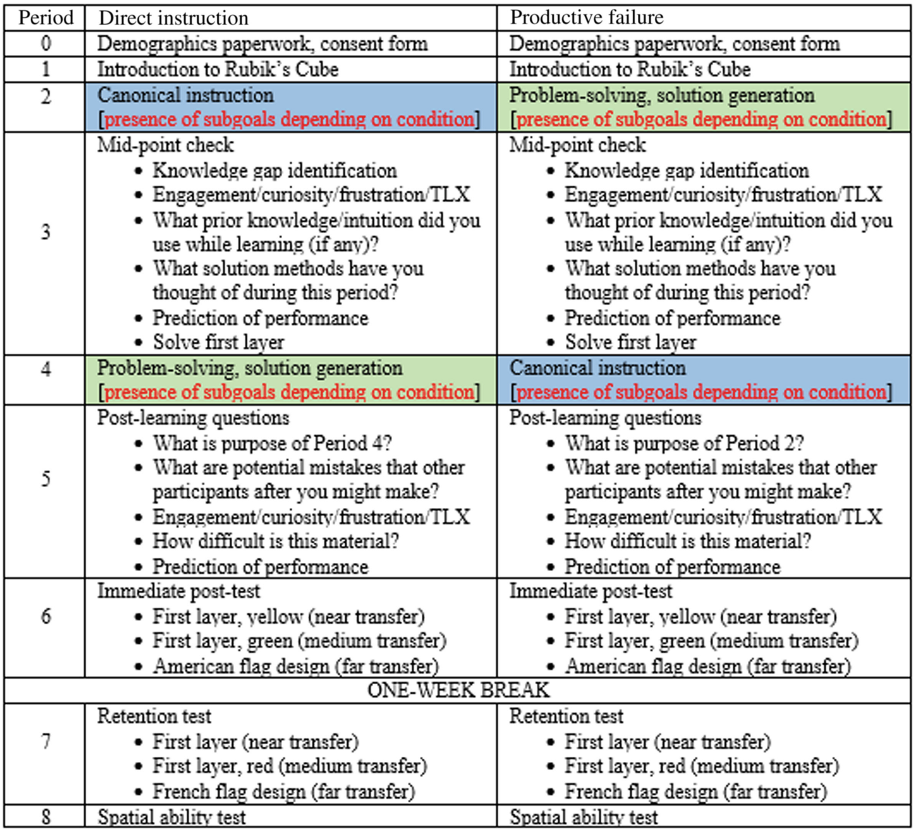


Period	Direct instruction	Productive failure
0	Demographics paperwork, consent form	Demographics paperwork, consent form
1	Introduction to cryptarithmetic	Introduction to cryptarithmetic
2	Canonical instruction [presence of subgoals depending on condition]	Problem-solving, solution generation [presence of subgoals depending on condition]
3	Mid-point check • Knowledge gap identification • Engagement/curiosity/frustration/TLX • What prior knowledge/intuition did you use while learning (if any)? • What solution methods have you thought of during this period? • Prediction of performance • Solve cryptarithmetic addition problem	Mid-point check • Knowledge gap identification • Engagement/curiosity/frustration/TLX • What prior knowledge/intuition did you use while learning (if any)? • What solution methods have you thought of during this period? • Prediction of performance • Solve cryptarithmetic addition problem
4	Problem-solving, solution generation [presence of subgoals depending on condition]	Canonical instruction [presence of subgoals depending on condition]
5	Post-learning questions • What is purpose of Period 4? • What are potential mistakes that other participants after you might make? • Engagement/curiosity/frustration/TLX • How difficult is this material? • Prediction of performance	Post-learning questions • What is purpose of Period 2? • What are potential mistakes that other participants after you might make? • Engagement/curiosity/frustration/TLX • How difficult is this material? • Prediction of performance
6	Immediate post-test • Addition, 2 numbers (near transfer) • Addition, 3 numbers (medium transfer) • Subtraction (far transfer)	Immediate post-test • Addition, 2 numbers (near transfer) • Addition, 3 numbers (medium transfer) • Subtraction (far transfer)
ONE-WEEK BREAK		
7	Retention test • Addition, 2 numbers (near transfer) • Addition, 3 numbers (medium transfer) • Subtraction (far transfer)	Retention test • Addition, 2 numbers (near transfer) • Addition, 3 numbers (medium transfer) • Subtraction (far transfer)
8	Logic/algebra ability test	Logic/algebra ability test

occurred. This expectation was realized in the above results. Many of the hypothesized advantages of PF methods were expected to instead become manifest during medium- and far-transfer problems, as well as retention problems, while DI methods' usage of isomorphic problems as practice (Clark et al. 2012) are conducive to performance on test problems that are similar to the practiced ones. The "regurgitative" nature of completing procedurally-similar problems immediately after learning increases the importance of streamlined problem-solving search processes often emphasized in DI while rendering the potentially deeper structural learning in PF relatively less useful.

However, a reason that DI was not hypothesized to actually overtake PF in immediate near-transfer task performance is that PF participants tend to report greater curiosity during canonical instruction than DI participants do (Loibl and Rummel 2014b), a phenomenon that was indirectly observed in this study when participants were surveyed about the purpose of the problem-solving learning period. In the cryptarithmetic domain, PF participants (M = 95%) were significantly more likely than DI participants (M = 24%) to say that the problem-solving period was to be used for

Table 2. Summary of Experiment 2 procedures (Rubik's Cube)

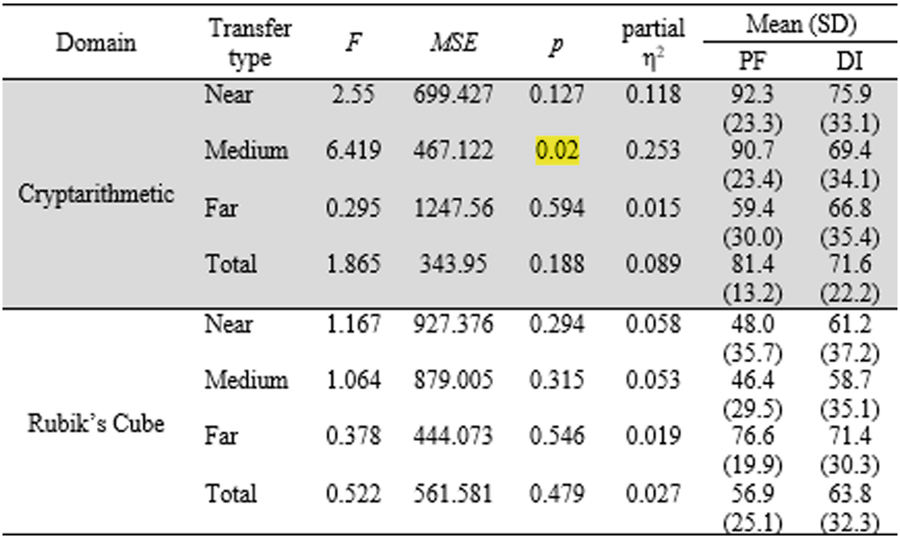


Period	Direct instruction	Productive failure
0	Demographics paperwork, consent form	Demographics paperwork, consent form
1	Introduction to Rubik's Cube	Introduction to Rubik's Cube
2	Canonical instruction [presence of subgoals depending on condition]	Problem-solving, solution generation [presence of subgoals depending on condition]
3	Mid-point check • Knowledge gap identification • Engagement/curiosity/frustration/TLX • What prior knowledge/intuition did you use while learning (if any)? • What solution methods have you thought of during this period? • Prediction of performance • Solve first layer	Mid-point check • Knowledge gap identification • Engagement/curiosity/frustration/TLX • What prior knowledge/intuition did you use while learning (if any)? • What solution methods have you thought of during this period? • Prediction of performance • Solve first layer
4	Problem-solving, solution generation [presence of subgoals depending on condition]	Canonical instruction [presence of subgoals depending on condition]
5	Post-learning questions • What is purpose of Period 4? • What are potential mistakes that other participants after you might make? • Engagement/curiosity/frustration/TLX • How difficult is this material? • Prediction of performance	Post-learning questions • What is purpose of Period 2? • What are potential mistakes that other participants after you might make? • Engagement/curiosity/frustration/TLX • How difficult is this material? • Prediction of performance
6	Immediate post-test • First layer, yellow (near transfer) • First layer, green (medium transfer) • American flag design (far transfer)	Immediate post-test • First layer, yellow (near transfer) • First layer, green (medium transfer) • American flag design (far transfer)
ONE-WEEK BREAK		
7	Retention test • First layer (near transfer) • First layer, red (medium transfer) • French flag design (far transfer)	Retention test • First layer (near transfer) • First layer, red (medium transfer) • French flag design (far transfer)
8	Spatial ability test	Spatial ability test

exploration (as opposed to practice and application), $F(1, 43) = 43.711$, $MSE = 0.128$, $p = 0.000$, partial $\eta^2 = 0.504$ (mean difference = 71%); a similar pattern of results for PF ($M = 100\%$) and DI ($M = 30.8\%$) held in the cube domain, $F(1, 49) = 54.044$, $MSE = 0.113$, $p = 0.000$, partial $\eta^2 = 0.524$ (mean difference = 69.1%). This question served to illuminate the mindsets of participants in the two instructional conditions and indeed revealed the exploratory approaches that PF participants tended to take.

According to Loibl and Rummel (2014b), initial unguided problem-solving periods in PF help learners to identify knowledge gaps that they are then more curious about resolving later when canonical instructions are presented; DI learners are not given intrinsic reason to pay as much attention to the canonical instructions. The benefits of the extra attention paid by PF participants to canonical instructions should be particularly evident during near-transfer test problems, given that the instructions focus on those types of problems. Moreover, not only were PF learners expected to be more curious and engaged, they were also expected to be more able to appreciate critical features of the presented canonical solutions due to comparisons of the strengths and

Table 3. Post-test score differences between instruction types



Domain	Transfer type	*F*	*MSE*	*p*	partial η^2	Mean (SD)	
						PF	DI
Cryptarithmetic	Near	2.55	699.427	0.127	0.118	92.3 (23.3)	75.9 (33.1)
	Medium	6.419	467.122	0.02	0.253	90.7 (23.4)	69.4 (34.1)
	Far	0.295	1247.56	0.594	0.015	59.4 (30.0)	66.8 (35.4)
	Total	1.865	343.95	0.188	0.089	81.4 (13.2)	71.6 (22.2)
Rubik's Cube	Near	1.167	927.376	0.294	0.058	48.0 (35.7)	61.2 (37.2)
	Medium	1.064	879.005	0.315	0.053	46.4 (29.5)	58.7 (35.1)
	Far	0.378	444.073	0.546	0.019	76.6 (19.9)	71.4 (30.3)
	Total	0.522	561.581	0.479	0.027	56.9 (25.1)	63.8 (32.3)

weaknesses of their invented solutions and the canonical ones (Moore and Schwartz 1998). Therefore, the advantages for each method were expected to "cancel out" to some extent, and the non-significant differences between PF and DI in both domains fulfilled those expectations.

Productive failure was hypothesized to produce significantly better performance in medium- and far-transfer problems, but that largely turned out not to be the case. The hypothesis was based on the notion that PF methods, just through the order of instruction, would require learners to combine heuristics and formal knowledge in ways that the "canonical instruction, then application practice" order in DI does not (Kapur and Bielaczyc 2011). This combining of various knowledge bases in PF was expected to provide learners with the resources to generate relatively wide ranges of solution methods (diSessa and Sherin 2000) due in part to the exploratory information gleaned from the initial problem-solving periods, and these different solution methods should have enabled better attempts at transfer problems that cannot be solved solely using canonical instructions. Participants in PF conditions ($M = 0.594$ unique solution strategies, $SD = 0.837$) did indeed attempt unique solution strategies more often than DI participants ($M = 0.219$, $SD = 0.420$) in cryptarithmetic, $F(1, 62) = 5.131$, $MSE = 0.439$, $p = 0.027$, partial $\eta^2 = 0.076$ (mean difference = 0.375), and the Rubik's Cube domain revealed similar differences between PF ($M = 0.781$, $SD = 0.552$) and DI ($M = 0.375$, $SD = 0.492$), $F(1, 62) = 9.648$, $MSE = 0.274$, $p = 0.003$, partial $\eta^2 = 0.135$ (mean difference = 0.406).

However, the use of unique strategies (those that were not explicitly explained in instructional material) apparently did not aid participants on tasks of medium and far transfer. While it still might be the case that those tasks do require novel and creative solution methods, perhaps the participants' invented methods were either not particularly relevant or did not enable the participants to learn deep structural information about the domain. Furthermore, deciphering the parts of a solution attempt that are generalizable, and those that are context-specific and ungeneralizable, is often difficult for novices due to a lack of experience (Patel et al. 1993), an issue that is likely magnified in PF when participants initially are relying more on their own heuristics to make assumptions about the domain.

4.2 Instruction Type Main Effects (Retention Assessments)

To analyze the retention test performance dependent measure, pre-existing ability (covariate), immediate post-test score (covariate), and the independent variables were used as predictors in a GLM. No significant retention score differences were found between PF ($M = 45.94\%$, $SD = 21.62\%$) and DI ($M = 48.62\%$, $SD = 20.10\%$) in cryptarithmetic, $F(1, 18) = 0.114$, $MSE = 376.147$, $p = 0.739$, partial $\eta^2 = 0.006$ (mean difference = 2.68%), and no significant retention score differences were found between PF ($M = 63.72\%$, $SD = 26.6\%$) and DI ($M = 66.35\%$, $SD = 26.6\%$), in Rubik's Cube, $F(1, 16) = 0.219$, $MSE = 171.214$, $p = 0.646$, partial $\eta^2 = 0.014$ (mean difference = 2.63%).

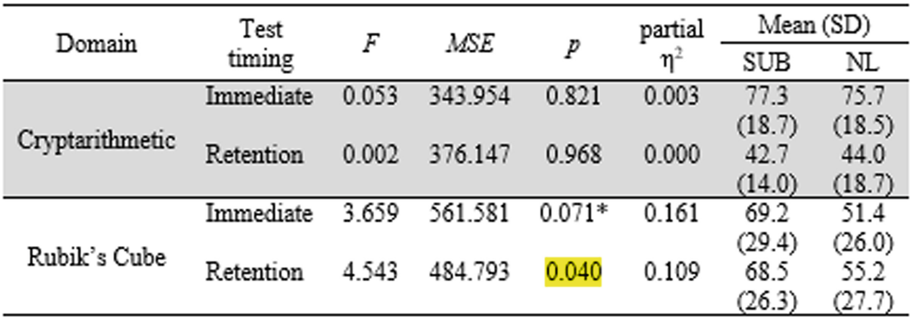
It was hypothesized that the inherently frequent activation of prior and long-term knowledge during initial PF problem-solving would require learners to connect new material with relatively stable information that they already knew (Kapur 2008) and furthermore lead to deeper encoding and assembling of schemas (Hiebert and Grouws 2007). As a result, the learning that ensued was expected to be more enduring and less fleeting, a difference that would be most apparent on retention problems. However, when surveyed on a Likert scale (1–7, 7 = most), participants in PF ($M = 4.25$, $SD = 2.11$) did not report using significantly more prior knowledge than DI ($M = 4.03$, $SD = 1.56$) in cryptarithmetic, $F(1, 62) = 0.223$, $MSE = 3.435$, $p = 0.639$, partial $\eta^2 = 0.004$ (mean difference = 0.22) and the differences between PF ($M = 3.31$, $SD = 1.79$) and DI ($M = 3.13$, $SD = 1.66$) were also not statistically significant in Rubik's Cube, $F(1, 62) = 0.189$, $MSE = 2.974$, $p = 0.665$, partial $\eta^2 = 0.003$ (mean difference = 0.188). For now, these data can inform some discussion and conclusions, but more-detailed analyses are likely needed in the future to examine, more generally, the differences in how PF and DI participants used problem-solving periods. Question prompts during problem-solving, for example, could enable researchers to more deeply study why a participant invented a particular solution strategy and whether that strategy contributed any generalizable domain knowledge through its use, or how a participant could be encouraged to activate more relevant prior and long-term knowledge.

In the current experiments, given that PF methods did not prove superior to DI in terms of encouraging participants to lean more on their prior knowledge, it is then unsurprising that retention performance was about equal between the two conditions. This pattern of findings on retention performance contradicts what "desirable difficulties" research would predict (i.e., slow performance improvements early on due to difficulty designed into the instruction, but better performance later; e.g., Bjork and Bjork 2011). It was expected that PF participants surpass their DI counterparts on assessments like the retention test, which was administered one week after the material was learned. Participants' struggles during the PF generation period would require deeper and more durable processing to navigate (i.e., connected to prior knowledge and/or self-generated heuristics), while DI participants would be more likely to fall into a false sense of competency because the learning process is relatively easier and performance on immediate tasks improves relatively quickly (Marsh and Butler 2013). However, survey measures such as workload (via NASA TLX) revealed that PF was not an appreciably more difficult experience than DI, and in some instances was actually reported to be an easier experience. Furthermore, not all participants in PF actually failed after the initial "struggle" period, which likely means that the given tasks were not difficult enough to yield productive failures and the associated benefits: 8 of 32 cryptarithmetic participants scored 100% on the mid-point check, while 6 of 32 Rubik's Cube participants performed likewise. Therefore, PF did not create enough desirable difficulty for participants.

4.3 Subgoal Label Main Effects (Immediate Post-test and Retention Assessments)

Upon examining the subgoal predictor of the GLMs for immediate test and retention test performance, a pattern emerged regarding scores across domains. Table 4 summarizes the scores of participants who received subgoals (SUB) and those who received non-labeled (NL) instructions (maximum possible test score is 100%):

Table 4. Test performance with subgoal- (SUB) and non-labeled (NL) instructions



Domain	Test timing	*F*	*MSE*	*p*	partial η^2	Mean (SD)	
						SUB	NL
Cryptarithmetic	Immediate	0.053	343.954	0.821	0.003	77.3 (18.7)	75.7 (18.5)
	Retention	0.002	376.147	0.968	0.000	42.7 (14.0)	44.0 (18.7)
Rubik's Cube	Immediate	3.659	561.581	0.071*	0.161	69.2 (29.4)	51.4 (26.0)
	Retention	4.543	484.793	0.040	0.109	68.5 (26.3)	55.2 (27.7)

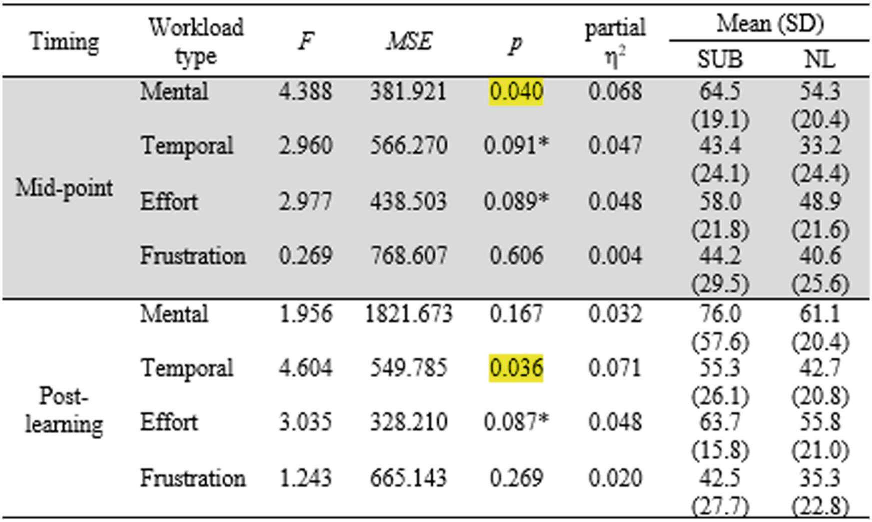
In the cryptarithmetic domain, subgoal labels appeared to make very little difference in test scores. Previous research has demonstrated that subgoal labels outline high-level information that can help learners organize domain content in meaningful ways (Atkinson et al. 2000), which theoretically should improve performance. However, it is probable that the college-educated participants did not require subgoal labels to help them organize content in a domain that is similar to algebra.

According to the data, Rubik's Cube participants were aided greatly by subgoal labels. Sweller (2010) notes that subgoals enable learners to focus just on fundamental structures of problems and not incidental features. In a domain like the Rubik's Cube in which participants likely do not possess much relevant experience, this generalizable information from subgoal labels is crucial so that participants do not extrapolate from concepts that might have been specific only to a given example.

4.4 Subgoal Label Main Effects (Workload)

Some evidence suggests that the subgoal labels in cryptarithmetic, if anything, served only to increase participant workload, possibly because of extra effort needed to interact with them. Tables 5 and 6 outline the workload data for both domains (maximum possible reported workload is 100%).

Table 5. Cryptarithmetic: Workload differences between SUB and NL instructions



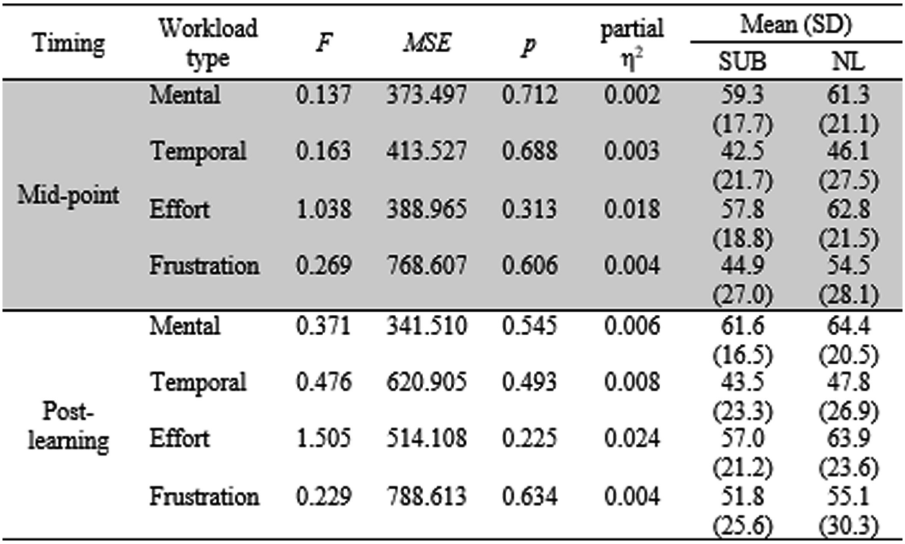
Timing	Workload type	*F*	*MSE*	*p*	partial η^2	Mean (SD)	
						SUB	NL
Mid-point	Mental	4.388	381.921	0.040	0.068	64.5 (19.1)	54.3 (20.4)
	Temporal	2.960	566.270	0.091*	0.047	43.4 (24.1)	33.2 (24.4)
	Effort	2.977	438.503	0.089*	0.048	58.0 (21.8)	48.9 (21.6)
	Frustration	0.269	768.607	0.606	0.004	44.2 (29.5)	40.6 (25.6)
Post-learning	Mental	1.956	1821.673	0.167	0.032	76.0 (57.6)	61.1 (20.4)
	Temporal	4.604	549.785	0.036	0.071	55.3 (26.1)	42.7 (20.8)
	Effort	3.035	328.210	0.087*	0.048	63.7 (15.8)	55.8 (21.0)
	Frustration	1.243	665.143	0.269	0.020	42.5 (27.7)	35.3 (22.8)

According to Table 5, subgoals increased workload significantly in the cryptarithmetic domain. Furthermore, subgoal labels did not improve performance in cryptarithmetic, suggesting that the increased load might have been extraneous. As was stated before, it is perhaps the case that subgoal labels were not necessary in the cryptarithmetic domain due to participants' familiarity with algebra, which could explain why participants reported subgoals as relatively taxing to interact with.

Subgoals did not increase workload in the Rubik's Cube domain, as demonstrated in Table 6. The participants likely found the Rubik's Cube subgoal labels to be essential information and therefore did not perceive them as difficult to engage. After all, the subgoal labels improved Rubik's Cube performance substantially.

Given the relatively robust findings in previous research regarding how subgoals reduce cognitive load in learners (e.g., Morrison et al. 2015), the findings in the current experiments are surprising. In future experiments, methods of implementing subgoal

Table 6. Rubik's Cube: Workload differences between SUB and NL instructions



Timing	Workload type	*F*	*MSE*	*p*	partial η^2	Mean (SD)	
						SUB	NL
Mid-point	Mental	0.137	373.497	0.712	0.002	59.3 (17.7)	61.3 (21.1)
	Temporal	0.163	413.527	0.688	0.003	42.5 (21.7)	46.1 (27.5)
	Effort	1.038	388.965	0.313	0.018	57.8 (18.8)	62.8 (21.5)
	Frustration	0.269	768.607	0.606	0.004	44.9 (27.0)	54.5 (28.1)
Post-learning	Mental	0.371	341.510	0.545	0.006	61.6 (16.5)	64.4 (20.5)
	Temporal	0.476	620.905	0.493	0.008	43.5 (23.3)	47.8 (26.9)
	Effort	1.505	514.108	0.225	0.024	57.0 (21.2)	63.9 (23.6)
	Frustration	0.229	788.613	0.634	0.004	51.8 (25.6)	55.1 (30.3)

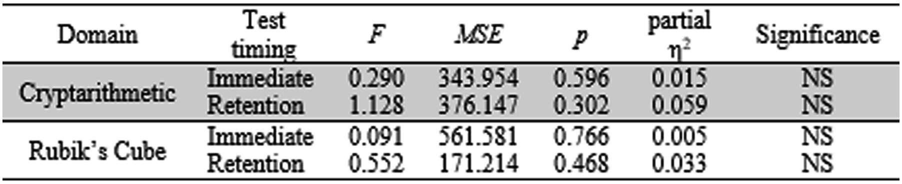
labels (e.g., frequency of labeling, type of content conveyed, learner role in generation of labels) could be manipulated to examine whether workload and performance results depend on the method of labeling.

4.5 Interaction Between Instruction Type and Presence of Subgoal Labels

Before the experiments started, subgoal labels presented during the PF generation period were expected to mitigate the chances that learners aimlessly pursued irrelevant objectives and formed structural misconceptions, risks inherent in any minimally-guided method (Brown and Campione 1994). While subgoal labels are generally important in DI materials as well, they were expected to be relatively less so because DI participants received instruction at the start of the learning process that was at least somewhat organized whether subgoals were labeled or not, and the participants were merely applying learned knowledge during the problem-solving phase, likely using the subgoal labels just as reminders. The data suggested that no such interaction between instruction type and subgoal labeling occurred, regardless of domain or timing of test (see Table 7).

Instead, a plausible explanation is that the positive effects of subgoals are robust across various methods of instruction, but not necessarily across all domains (per previous findings). After all, the key purpose of subgoal labels is helping learners recognize fundamental components of a domain (Catrambone 1998), a useful aid regardless of whether a learner is using productive failure or direct instruction. However, the extent to which that aid increases performance might depend on domain

Table 7. Interaction between instruction type and subgoal labeling (test scores)



Domain	Test timing	*F*	*MSE*	*p*	partial η^2	Significance
Cryptarithmetic	Immediate	0.290	343.954	0.596	0.015	NS
	Retention	1.128	376.147	0.302	0.059	NS
Rubik's Cube	Immediate	0.091	561.581	0.766	0.005	NS
	Retention	0.552	171.214	0.468	0.033	NS

familiarity and how easily learners can discern fundamental components on their own in that given domain.

In summary, subgoal labels improved performance in the Rubik's Cube domain, regardless of instruction type, but failed to improve performance in the cryptarithmetic domain (also regardless of instruction type). A potential future research direction could involve manipulating the scaffolding mechanism used in PF instruction to examine whether other scaffolding mechanisms are more reliable across domains (e.g., self-explanation prompts, social discourse). Preventing learners from failing unproductively and veering too far off track is a scaffolding mechanism that has been shown to be effective in general (e.g., training wheels; Carroll and Carrithers 1984), but other methods could prove superior in particular learning contexts. A systematic examination of domains is also necessary to study how these various scaffolding mechanisms interact with domains of particular characteristics; for example, the motivational aspects of group discourse (Lin et al. 1999) could improve learning relatively substantially in inherently uninteresting domains, but not spur much improvement in domains that are inherently more interesting.

5 Conclusions

In general, PF methods in the present studies produced some ostensibly positive ancillary developments for learners (exploratory mindsets, diverse solution attempts, and occasionally lower workload). However, those ancillary developments did not lead to the ultimate goal of increasing post-test and retention test performance. This phenomenon suggests questions for further study such as whether the relevance and quality of learners' solution attempts should be regulated somehow (perhaps through the use of scaffolding methods other than subgoals), or whether lower workload is inherently beneficial.

Research in productive failure is still in its early stages and therefore much work remains to be done in improving the method itself. Potential improvements include explicit elicitation of prior domain knowledge, more meaningful subgoal labels, and group learning implementation. Replicating findings in various domains will also be an important task, given that people have access to learning wider varieties of information than ever but most learning research still centers on just science- and mathematics-related domains. Some patterns of results from the current experiments changed depending on domain, but systematic selection of domains would enable researchers to find more precisely the dimensions and characteristics of domains that drive changes in results.

References

Anthony, W.S.: Learning to discover rules by discovery. J. Educ. Psychol. **64**(3), 325–328 (1973)

Atkinson, R.K., Derry, S., Renkl, A., Wortham, D.: Learning from examples: instructional principles from the worked examples research. Rev. Educ. Res. **70**(2), 181–214 (2000)

Bjork, E.L., Bjork, R.A.: Making things hard on yourself, but in a good way: creating desirable difficulties to enhance learning. In: Gernsbacher, M.A., et al. (eds.) Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society. Worth Publishers, New York (2011)

Blodgett, H.C.: The effect of the introduction of reward upon the maze performance of rats. Univ. Calif. Publ. Psychol. **4**, 113–134 (1929)

Bransford, J.D., Schwartz, D.L.: Rethinking transfer: a simple proposal with multiple implications. In: Iran-Nejad, A., Pearson, P.D. (eds.) Review of Research in Education. American Educational Research Association, Washington, DC, pp. 61–101 (1999)

Brown, A., Campione, J.: Guided discovery in a community of learners. In: McGilly, K. (ed.) Classroom Lessons: Integrating Cognitive Theory and Classroom Practice, pp. 229–270. MIT Press, Cambridge (1994)

Carrier, M., Pashler, H.: The influence of retrieval on retention. Mem. Cogn. **20**, 633–642 (1992)

Carroll, J.M., Carrithers, C.: Training wheels in a user interface. Commun. ACM **27**(8), 800–806 (1984)

Catrambone, R.: The subgoal learning model: creating better examples so that students can solve novel problems. J. Exp. Psychol. Gen. **127**(4), 355–376 (1998)

Chen, D.: The Role of Struggle and Productive Failure in Learner Assistance (2016, Unpublished manuscript)

Chi, M.T.H.: Self-explaining: the dual processes of generating inferences and repairing mental models. In: Glaser, R. (ed.) Advances in Instructional Psychology, pp. 161–238. Lawrence Erlbaum, Mahwah (2000)

Clark, R.E., Kirschner, P.A., Sweller, J.: Putting students on the path to learning: the case for fully guided instruction. Am. Educ. **36**, 6–11 (2012)

Clement, J.: Non-formal reasoning in science: the use of analogies, extreme cases, and physical intuition. In: Voss, J.F., Perkins, D.N., Siegel, J. (eds.) Informal Reasoning and Education. Lawrence Erlbaum Associates, Hillsdale (1991)

Cope, P., Simmons, M.: Some effects of limited feedback on performance and problem-solving strategy in a logo microworld. J. Educ. Psychol. **86**(3), 368–379 (1994)

diSessa, A., Hammer, D., Sherin, B., Kolpakowski, T.: Inventing graphing: children's meta-representational expertise. J. Math. Behav. **10**(2), 117–160 (1991)

diSessa, A., Sherin, B.L.: Meta-representation: an introduction. J. Math. Behav. **19**, 385–398 (2000)

Fiore, S.M., Scielzo, S., Jentsch, F., Howard, M.L.: Understanding performance and cognitive efficiency when training for x-ray security screening. In: Proceedings of the HFES 50th Annual Meeting, pp. 2610–2614. HFES, Santa Monica (2006)

Gorman, J., Cooke, N., Amazeen, P.: Training adaptive teams. Hum. Factors **52**, 295–307 (2010)

Hiebert, J., Grouws, D.: The effects of classroom mathematics teaching on students' learning. In: Lester, F.K. (ed.) 2nd Handbook of Research on Mathematics Teaching and Learning, pp. 371–404. Information Age, Charlotte (2007)

Jonassen, D.: Objectivism vs. constructivism. Educ. Tech. Res. and Dev. **39**(3), 5–14 (1991)

Kulhavy, R.W., Stock, W.A.: Feedback in written instruction: the place of response certitude. Educ. Psychol. Rev. **1**(4), 279–307 (1989)

Kapur, M.: Productive failure. Cogn. Instr. **26**(3), 379–424 (2008)

Kapur, M.: A further study of productive failure in mathematical problem solving: unpacking the design components. Instr. Sci. **39**, 561–579 (2011)

Kapur, M., Bielaczyc, K.: Classroom-based experiments in productive failure. In: Carlson, L., Holscher, C., Shipley, T. (eds.) Proceedings of the 33rd annual conference of the Cognitive Science Society, pp. 2812–2817. Cognitive Science Society, Austin (2011)

Kapur, M., Dickson, L., Yhing, T.P.: Productive failure in mathematical problem solving. Instr. Sci. **38**(6), 523–550 (2010)

Kapur, M., Lee, K.: Designing for productive failure in mathematical problem solving. In: Proceedings of the 31st Annual Conference of the Cognitive Science Society, pp. 2632–2637. Cognitive Science Society, Austin (2009)

Kerr, R., Booth, B.: Specific and varied practice of a motor skill. Percept. Mot. Ski. **46**, 395–401 (1978)

Kirschner, P.A., Sweller, J., Clark, R.E.: Why minimal guidance during instruction does not work: an analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based learning. Educ. Psychol. **41**(2), 75–86 (2006)

Koedinger, K.R., Aleven, C.: Exploring the assistance dilemma in experiments with cognitive tutors. Educ. Psychol. Rev. **19**, 239–264 (2007)

Kornell, N., Bjork, R.A.: A stability bias in human memory: overestimating remembering and underestimating learning. J. Expermintal. Psychol.: Gen. **138**, 449–468 (2009)

Lin, X., Hmelo, X., Kinzer, C.K., Secules, T.J.: Designing technology to support reflection. Educ. Technol. Res. Dev. **47**(3), 43–62 (1999)

Loibl, K., Rummel, N.: The impact of guidance during problem-solving prior to instruction on students' inventions and learning outcomes. Instr. Sci. **42**, 305–326 (2014a)

Loibl, K., Rummel, N.: Knowing what you don't know makes failure productive. Learn. Instr. **34**, 74–85 (2014b)

Marsh, E.J., Butler, A.C.: Memory in educational settings. In: Resiberg, D. (ed.) The Oxford Handbook of Cognitive Psychology, pp. 299–317. Oxford University Press (2013)

Mathan, S., Koedinger, K.R.: Recasting the feedback debate: benefits of tutoring error detection and correction skills. In: Hoppe, H.U., et al. (eds.) Artificial Intelligence in Education, pp. 13–20. IOS Press (2003)

Moore, J.L., Schwartz, D.L.: On learning the relationship between quantitative properties and symbolic representations. In: Proceedings of the International Conference of the Learning Sciences, pp. 209–214. Erlbaum, Mahwah (1998)

Morrison, B.B., Margulieux, L.E., Guzdial, M.: Subgoals, context, and worked examples in learning computing problem solving. In: Proceedings of the 11th Annual International Conference on International Computing Education Research, pp. 21–29. ACM, New York (2015)

Patel, V.L., Groen, G.J., Norman, G.R.: Reasoning and instruction in medical curricula. Cogn. Instr. **10**, 335–378 (1993)

Rourke, A., Sweller, J.: The worked-example effect using ill-defined problems: learning to recognize designers' styles. Learn. Instr. **19**, 185–199 (2009)

Schank, R.: Virtual Learning: A Revolutionary Approach to Building a Highly-Skilled Workforce. McGraw-Hill, New York (1997)

Schmidt, R.A., Bjork, R.A.: New conceptualizations of practice: common principles in three paradigms suggest new concepts for training. Psychol. Sci. **3**(4), 207–217 (1992)

Schwartz, D.L., Martin, T.: Inventing to prepare for future learning: the hidden efficiency of encouraging original student production in statistics instruction. Cogn. Instr. **22**(2), 129–184 (2004)

Shea, J.B., Morgan, R.L.: Contextual interference effects on the acquisition, retention, and transfer of a motor skill. J. Exp. Psychol.: Hum. Learn. Mem. **5**, 179–187 (1979)

Siegler, R.S.: Microgenetic studies of self-explanation. In: Garnott, N., Parziale, J. (eds.) Microdevelopment: A Process-Oriented Perspective for Studying Development and Learning, pp. 31–58. Cambridge University Press, Cambridge (2002)

Soderstrom, N.C., Bjork, R.A.: Learning versus performance: an integrative review. Perspect. Psychol. Sci. **10**(2), 176–199 (2015)

Sweller, J.: Element interactivity and intrinsic, extraneous, and germane cognitive load. Educ. Psychol. Rev. **22**(2), 123–138 (2010)

Sweller, J., Mawer, R., Howe, W.: The consequences of history-cued and means-ends strategies in problems solving. Am. J. Psychol. **95**, 455–484 (1982)

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

VanLehn, K., Siler, S., Murray, C., Yamauchi, T., Baggett, W.B.: Why do only some events cause learning during human tutoring? Cogn. Instr. **21**(3), 209–249 (2003)

Wineburg, S.S., Foumier, J.E.: Contextualized thinking in history. In: Carretero, M., Voss, J.F. (eds.) Cognitive and Instructional Processes in History and the Social Sciences, pp. 285–308. Erlbaum, Hillsdale (1994)

Young, M.D., Healy, A.F., Gonzalez, C., Dutt, V., Bourne, L.E.: Effects of training with added difficulties on RADAR detection. Appl. Cogn. Psychol. **25**, 395–407 (2011)