COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 38 (2014) 1008–1022

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12107

Productive Failure in Learning Math

Manu Kapur

National Institute of Education, Nanyang Technological University

Received 27 July 2012; received in revised form 15 April 2013; accepted 15 May 2013

Abstract

When learning a new math concept, should learners be first taught the concept and its associated procedures and then solve problems, or solve problems first even if it leads to failure and then be taught the concept and the procedures? Two randomized-controlled studies found that both methods lead to high levels of procedural knowledge. However, students who engaged in problem solving before being taught demonstrated significantly greater conceptual understanding and ability to transfer to novel problems than those who were taught first. The second study further showed that when given an opportunity to learn from the failed problem-solving attempts of their peers, students outperformed those who were taught first, but not those who engaged in problem solving first. Process findings showed that the number of student-generated solutions significantly predicted learning outcomes. These results challenge the conventional practice of direct instruction to teach new math concepts and procedures, and propose the possibility of learning from one's own failed problem-solving attempts or those of others before receiving instruction as alternatives for better math learning.

Keywords: Mathematics education; Learning; Problem solving; Productive failure; Vicarious failure

1. Introduction

Suppose one wants to teach students a math concept (and its associated procedures) that is novel to them, say standard deviation (SD). The traditional, most prevailing method is to first teach students the concept and procedures of SD and then get them to solve problems requiring those concept and procedures. This sequence of instruction followed by problem solving is commonly known as *direct instruction* (DI; Kirschner, Sweller, & Clark, 2006). A contrasting method is one that reverses the sequence, that is, engages students in problem solving first and then teaches them the concept and procedures. I call this sequence of problem solving followed by instruction *productive failure* (PF; Kapur,

Correspondence should be sent to Manu Kapur, National Institute of Education, Nanyang Technological University, 1 Nanyang Walk, Singapore 637616. E-mail: manu.kapur@nie.edu.sg

2010, 2012). By failure, I simply mean that students will typically *fail* to generate or discover the correct solution(s) by themselves. However, to the extent that students are able to use their prior knowledge to generate suboptimal or even incorrect solutions to the problem, the process can be productive in preparing them to learn better from the subsequent instruction that follows (Kapur & Bielaczyc, 2012; Schwartz & Martin, 2004). In this way, PF combines the benefits of exploratory problem solving and instruction, thereby mitigating the possibility that students do not discover the correct concepts and procedures on their own (Kapur & Rummel, 2012).

There are several reasons to believe in the effectiveness of DI. First, it affords students the opportunities to attend to and acquire the correct procedures and knowledge, while concomitantly reducing the probability of encoding of errors and misconceptions (Sweller & Chandler, 1991). Without such instruction, students may not be able to discover correct knowledge and procedures on their own (Klahr & Nigam, 2004). Second, when students do not have the knowledge to solve a problem, they often search the problem space for solutions by engaging in resource-intensive processes such as trial and error or means-ends analysis, which burden the limited working memory capacity. Because all conscious processing happens in the working memory, working memory is less available for learning new concepts and procedures if it is mainly occupied with such a search of the problem space (Kirschner et al., 2006). By showing the learner exactly what to do and how to do it, DI reduces this burden on the cognitive resources, thereby facilitating the development of correct domain knowledge and procedures (Klahr & Nigam, 2004; Sweller & Chandler, 1991). Finally, DI can also mitigate problems associated with learner disengagement and frustration that can arise in starting with problem solving first (Hardiman, Pollatsek, & Weil, 1986).

There are also several reasons to believe in the effectiveness of PF. First, starting with problem solving may be better at activating and differentiating relevant prior knowledge provided students are able to use their priors to generate suboptimal or even incorrect solutions to the problem (DeCaro & Rittle-Johnson, 2012; Schwartz, Chase, Oppezzo, & Chin, 2011). Even though generation of solutions may place higher cognitive demands and is difficult for novices, such difficulty can aid encoding and schema assembly (Hiebert & Grouws, 2007; Schmidt & Bjork, 1992), and prepare students to learn better from the subsequent instruction (Kapur, 2012; Schwartz & Martin, 2004). Second, to the extent that students are able to persist in problem solving in spite of the higher cognitive demands, generating solutions may afford students greater agency, and therefore engage them more (diSessa, Hammer, Sherin, & Kolpakowski, 1991). Third, generating solutions prior to instruction may also help students notice the inconsistencies and realize the limits of their prior knowledge (DeCaro & Rittle-Johnson, 2012). Fourth, prior knowledge activation and differentiation may afford greater opportunities for comparisons between student-generated solutions and correct solutions, thereby helping students' attend to and better encode critical features of the new concept. Consequently, such comparisons may increase the likelihood of students selecting correct knowledge and procedures over incorrect ones (Siegler, 1994, 2002).

The aim of this paper was to test the competing sets of hypotheses supporting DI and PF through two experimental studies. Study 1 compares the DI with PF for teaching the concept and procedures of SD. Study 2 extends Study 1 by providing a stricter comparison for PF in *vicarious failure* (VF). Although one can learn vicariously from both processes and outcomes, for the purposes of this study, I operationalize VF as learning vicariously from the outcomes, that is, the case where students study and evaluate their peers' solutions before receiving instruction.

2. Study 1: Comparing learning from PF and DI

2.1. Participants and design

Participants were 75 ninth-grade mathematics students (14–15 year olds; 38 boys, 37 girls) from a co-ed private school in the national capital region of India. All students were of Indian ethnicity. Students had not had any instruction on SD because this topic is not taught until the eleventh grade.

2.1.1. Pretest and math ability

One week before the study, all students took a paper-and-pencil pretest ($\alpha = 0.75$) comprising six items: three multiple-choice items on central tendencies, one multiple-choice item each on distributions and SD, and one open-response item on SD. Each correct answer was awarded one mark. The open-response item was scored as correct or incorrect by two independent raters with an inter-rater reliability of 1.00 as no student was able to answer it correctly. Composite score on the pretest was scaled linearly to 10; this score upon 10 formed the measure of prior knowledge. In addition, the school provided data from their most recent standardized test on mathematics for the cohort, which was taken as a measure of math ability.

2.1.2. Design

On the day of the experiment, students experienced two 1 h phases one after the other: a problem-solving phase and an instruction phase. The experimental manipulation was in the random assignment of students to the *sequence* in which they experienced the two phases. Although students came from different classes, they were randomly combined to form each condition. In PF condition ($n = 37$; 19 boys, 18 girls), students first experienced the problem-solving phase followed by the instruction phase. In the DI condition ($n = 38$; 19 boys; 19 girls), students experienced instruction phase followed by the problem-solving phase. The same teacher taught both the conditions. The time on task, the number of problems solved, and materials for each of the phases were identical in both the conditions. Neither the teacher nor the students were made aware of the experimental hypotheses being tested.

In the problem-solving phase, students were seated in a classroom and asked to generate as many solutions as possible to a problem on SD (Fig. 1). Students worked individually

Who is the most consistent Basketball player?

Mike and Dave are the top two players in a Basketball league. The table shows the number of points scored by Mike and Dave over the course of 20 games in the league.

An award has to be given to the more consistent player of the two. The decision has to be made mathematically.

Design as many measures of consistency as you can to determine the more consistent player.

Show all working.

Points scored by Mike and Dave

Game	Mike	Dave
1	24	23
2	19	19
3	24	26
4	20	24
5	25	20
6	21	21
7	25	23
8	21	24
9	26	25
10	22	29
11	26	24
12	22	22
13	27	25
14	23	24
15	27	27
16	23	23
17	28	24
18	24	28
19	29	24
20	24	25

Fig. 1. The problem given to students during the problem-solving phase.

without any help as indeed one would in an examination setting. Students were provided with blank A4 sheets of paper and were asked to clearly number and demarcate their solutions. Because students could only rely on their prior knowledge to generate solutions, the number of solutions generated by a student was taken as a proxy measure of his/her prior knowledge activation and differentiation (see supplementary materials Appendix A for coding and counting of student-generated solutions).

In the instruction phase, students were seated in a classroom and their teacher—an experienced mathematics teacher at the high-school level—taught the concept and procedures of SD. The teaching of SD was organized around four problems that included cycles of teacher modeling through worked-out examples demonstrating the concept and procedures, student practice, and feedback (see supplementary materials Appendix B). The design of the four problems in the form of simultaneously presented contrasting cases was done in line with the well-established finding that contrasting cases help students attend to critical features of the problem, and therefore aid learning (Rittle-Johnson & Star, 2009; Schwartz et al., 2011). No other problems or solutions (e.g., incorrect or suboptimal solutions) were used. Throughout this phase, the teacher directed attention to the critical features of SD and highlighted common errors and misconceptions. Student performance on the fourth problem was taken as an indicator of their learning of the procedure for calculating and conceptually interpreting SD. Two independent raters scored students solutions on the fourth problem as either incorrect or correct with an inter-rater reliability of 0.91. Solutions that deployed the correct formulation and procedure but contained computational errors were scored as correct.

2.1.3. Mental effort and engagement

Immediately after each phase, all students took a five-item, five-point (1 = strongly disagree to 5 = strongly agree) Likert scale engagement survey ($\alpha = 0.79$). Students also estimated their amount of mental effort using a nine-point rating scale that is commonly used in the cognitive load literature as a measure of cognitive load (Paas, 1992). Thus, each student had two engagement scores (calculated as the average rating of the five items) and two mental effort scores.

2.1.4. Posttest

Immediately after the second phase, all students took a 40-min posttest ($\alpha = 0.84$) comprising 19 items targeting: (a) procedural knowledge (testing the basic procedure for computing and interpreting SD; 2 multiple-choice items), (b) conceptual understanding (testing understanding of critical features of SD and deducing its mathematical properties; 11 multiple-choice and 3 open-response items), and (c) transfer (testing whether students can adapt knowledge of SD to solve problems on the concept of normalization not taught during instruction; 3 multiple-choice items). Each correct answer was awarded one mark. The open-response items for conceptual understanding were scored as correct or incorrect by two independent raters with an inter-rater reliability of 0.97. Composite scores for each type of item were scaled linearly to 10. This score upon 10 for the three types of items, namely procedural knowledge, conceptual understanding, and transfer, formed the three dependent variables.

All instruments (pretest, posttest, engagement survey, and mental effort rating) are available as supplementary materials online.

2.2. Results

2.2.1. Process results

2.2.1.1. Problem-solving phase: Productive failure students produced on average about six solutions, $M = 6.08$, $SD = 1.53$. Students used central tendencies (mean, median, and mode) as well as range. Students generated dot diagrams and line graphs to qualitatively examine the clustering and fluctuation trends. Another common solution was to count the frequency with which a player scored above, below, and at the mean to quantify the clustering at the mean. Students also calculated year-on-year deviation to argue that the greater the sum (or sometimes the average) of the deviations, the lower the consistency.

By comparison, DI students produced on average about three solutions, $M = 2.85$, $SD = 0.45$ during their problem-solving phase. All DI students generated the canonical solution, in addition to using the range, dot diagrams, or line graphs. More important, 100% of DI students were able to generate the canonical solution to the problem, whereas none of the PF students were able to do so.

2.2.1.2. Instruction phase: The percentages of PF and DI students with correct solutions on the fourth problem were 97.3% and 97.4%, respectively. Table 1 presents the descriptive

Table 1
Study 1: Summary of math ability, pretest, mental effort, engagement, and posttest performance

		Productive Failure		Direct Instruction	
	Max	*M*	*SD*	*M*	*SD*
Math ability	100	84.8	11.92	85.02	11.16
Prior knowledge (pretest)	10	4.65	1.70	4.53	2.11
Mental effort (PS)	9	7.65	0.86	6.18	0.87
Mental effort (I)	9	7.38	0.76	6.00	0.90
Engagement (PS)	5	4.46	0.41	4.39	0.48
Engagement (I)	5	4.48	0.44	4.48	0.49
Posttest					
Procedural knowledge	10	9.24	1.38	9.47	1.27
Conceptual understanding	10	6.33	1.25	3.84	1.24
Transfer	10	5.37	1.46	3.11	1.51

I, instruction phase; PS, problem-solving phase.

statistics for mental effort and engagement as well as summary of pre- and posttest performance.

2.2.1.3. Mental effort and engagement: Productive failure students reported significantly greater mental effort than DI students during the problem-solving phase ($F[1, 71] = 53.98$, $p < .001$, $d = 1.70$) as well as the instruction phase ($F[1, 71] = 52.09$, $p < .001$, $d = 1.66$). There were no significant differences on engagement scores between the conditions during the two phases.

2.2.2. Pre–posttest results

2.2.2.1. Math ability and prior knowledge: ANOVAs with experimental condition as the between-subjects factor revealed no significant difference between the two conditions on math ability ($F[1, 73] = 0.01$, $p = .933$), and prior knowledge ($F[1, 73] = 0.08$, $p = .784$).

2.2.2.2. Posttest: A MANCOVA with scores on procedural knowledge, conceptual understanding, and transfer as the three dependent variables, experimental condition as the between-subjects factor, and math ability, prior knowledge, average engagement score, and average mental effort score as the four covariates revealed significant multivariate main effects only of math ability ($F[3, 69] = 4.36$, $p = .007$) and condition ($F[3, 67] = 13.49$, $p < .001$). There were no other significant effects of the covariates or any interaction effects. The model was not sensitive to the exclusion of engagement and mental effort scores.

Univariate ANCOVAs suggested that there was no significant difference between PF and DI students on procedural knowledge ($F[1, 69] = 0.02$, $p = .896$). However, PF students significantly outperformed DI students on conceptual understanding ($F[1, 69] = 40.23$, $p < .001$, $d = 2.00$) and transfer ($F[1, 69] = 16.92$, $p < .001$, $d = 1.52$).

Finally, the number of solutions generated by PF students during their problem-solving phase was significantly correlated with their posttest scores on conceptual understanding ($r[37] = .65$, $p < .001$) and transfer ($r[37] = .81$, $p < .001$), but not with procedural knowledge. There were no such significant correlations in the DI condition.

2.3. Discussion

Productive failure students, in spite of reporting greater mental effort than DI students, significantly outperformed DI students on conceptual understanding and transfer without compromising procedural knowledge.[1] Evidence therefore supports the hypothesis that the PF method activated and differentiated students' prior knowledge during the problem-solving phase, which may have prepared them to learn from the subsequent instruction phase. The significant correlation between the number of solutions generated by PF students during the problem-solving phase—a proxy indicator of prior knowledge activation and differentiation—and their conceptual understanding and transfer performance on the posttest lends further credibility to this explanation.

However, Study 1 raises a further question: If prior knowledge activation and differentiation is essential for learning from subsequent instruction, then is it necessary to have students generate their own solutions or can they simply study and evaluate solutions generated by their peers before receiving instruction (that is, learn from VF)? Study 2 addresses this question by comparing learning from PF, VF, and DI.

Once again, there are competing sets of hypotheses supporting the case for PF and VF. On one hand, students who generate solutions may understand their own solutions better than students who study and evaluate them (Roll, Aleven, & Koedinger, 2011). Consequently, PF students may engage in deeper comparisons between the solutions than VF students, which may increase the likelihood of PF students attending to the critical features of the problem (Terwel, van Oers, van Dijk, & van den Eeden, 2009). As argued earlier, such comparisons may also increase the likelihood of PF students selecting correct procedures and features over incorrect ones (Durkin & Rittle-Johnson, 2012; Siegler, 2002). Finally, students who generate solutions may benefit from greater agency (diSessa et al., 1991) and therefore be more engaged than VF students in learning from the subsequent instruction.

On the other hand, studying and evaluating solutions presented as worked examples may take up fewer cognitive resources than generating solutions, which means that VF students may have more resources for better encoding and schema acquisition (Kirschner et al., 2006; Roll et al., 2011). An extensive body of empirical work suggests that learning from worked examples is significantly better than learning from problem solving (for a review, see Kirschner et al., 2006). Furthermore, research on expertise suggests that although both generation and evaluation require domain knowledge (Bransford, Brown, & Cocking, 2000), generation may place a greater burden on the learners' domain knowledge than evaluation (in addition to burdening the working memory). In fact, the expertise reversal effect (Kalyuga, Ayres, Chandler, & Sweller, 2003) suggests that students with more domain knowledge may benefit more from the cognitively more demanding

processes. As both PF and VF students do not have the necessary domain knowledge to solve the problem, it follows that VF students may benefit more from evaluation of solutions than PF students from generation of solutions.

Given that Study 1 demonstrated the effectiveness of PF over DI, and the main difference between PF and VF is one of generating solutions and generating evaluations of solutions, that is, both are preparatory activities prior to instruction aimed at activating and differentiating prior knowledge, it can be hypothesized that when compared to DI students, VF students will also likely benefit from prior knowledge activation and differentiation, preparing them to learn better than DI students from instruction.

3. Study 2: Comparing learning from PF, VF, and DI

3.1. Participants and design

Participants were 111 ninth-grade mathematics students (14–15 year olds; 55 boys, 56 girls) from the same school as in Experiment 1, but from a different cohort. A randomized-controlled, pre-post design was used to assign students to one of three conditions: DI (n = 38; 19 boys, 19 girls), PF (n = 37; 18 boys; 19 girls), or VF (n = 36; 18 boys, 18 girls).

The PF and DI conditions were exactly the same as in Study 1. The same teacher from Study 1 taught the instruction phase to all three conditions. The VF condition differed from the PF condition in only one aspect: The problem-solving phase was replaced with an evaluation phase in which students were given 1 hour to study and evaluate student-generated solutions (available from Study 1). Each solution was presented on an A4 sheet of paper with the prompt: "Evaluate whether this solution is a good measure of consistency. Explain and give reasons to support your evaluation." The number of solutions given was pegged to the average number of solutions produced by the PF groups, that is, six. The most frequently generated solutions by the PF students were chosen for VF condition. Because student-generated solutions often lacked clarity in their presentation that may make it difficult for other students to understand, let alone evaluate them, they were converted into well-designed worked examples. Figs. S1–S3 in the supplementary materials present the six solutions. Vicarious failure students received the solutions one-by-one counterbalanced for order and were given about 10 min for each. As in the PF condition, no help was provided during the evaluation phase.

3.1.1. Evaluation training

Pilot work suggested that while students did not have problems studying and understanding the solutions, students did better evaluations when they were given an example of what constituted a mathematically valid evaluation. Therefore, the first of the six student-generated solutions on central tendencies was provided as an example that contrasted a valid with an invalid evaluation (see Fig. S1). Students were given 10 min to go through this evaluation example after which they evaluated the remaining five solutions.

Two independent raters coded evaluations for each solution as valid or invalid using the coding scheme described in Appendix A of the supplementary materials. Table S1 in the supplementary materials presents examples of valid and invalid evaluation for each solution.

3.2. Results

3.2.1. Process results

3.2.1.1. Problem-solving phase for PF and DI students: Productive failure students produced on average about six solutions (M = 6.15, SD = 0.97). By comparison, DI students produced on average about three solutions (M = 2.95, SD = 0.39) during their problem-solving phase. Once again, 100% of DI students were able to generate the canonical solution to the problem, whereas none of the PF students were able to do so.

3.2.1.2. Evaluation phase for VF students: On average, VF students produced 0.85 (SD = 0.41) valid and 0.35 (SD = 0.23) invalid evaluations per solution. These results suggest that VF students were able to understand and evaluate the solutions.

3.2.1.3. Instruction phase for all students: The percentages of PF, VF, and DI students with correct solutions on the fourth problem were 94.7%, 94.7%, and 97.4%, respectively. Table 2 presents the descriptive statistics for mental effort and engagement as well as summary of pre- and posttest performance.

3.2.1.4. Mental effort and engagement: For the problem-solving phase, there was a significant multivariate effect of condition on mental effort scores (F[2, 106] = 10.69, $p < .001$). Planned pairwise comparisons showed that PF students reported significantly

Table 2
Study 2: Summary of math ability, pretest, mental effort, engagement, and posttest performance

		Productive Failure		Vicarious Failure		Direct Instruction	
	Max	*M*	*SD*	*M*	*SD*	*M*	*SD*
Math ability	100	85.0	12.37	85.4	13.06	84.46	12.15
Pretest	10	4.48	2.23	4.05	2.76	4.37	2.50
Mental effort (PS)	9	7.38	1.04	6.39	1.02	6.30	1.19
Mental effort (I)	9	7.30	1.02	6.44	1.03	5.82	1.11
Engagement (PS)	5	4.42	0.43	4.36	0.42	4.28	0.55
Engagement (I)	5	4.43	0.44	4.32	0.38	4.36	0.52
Posttest							
Procedural knowledge	10	8.86	1.90	8.89	1.85	9.21	1.85
Conceptual understanding	10	6.57	1.26	4.72	1.48	3.57	1.41
Transfer	10	5.36	1.47	3.31	1.87	3.06	2.09

I, instruction phase; PS, problem-solving phase.

greater mental effort than VF students ($F[1, 71] = 16.85$, $p < .001$, $d = 0.96$), as well as DI students ($F[1, 73] = 16.10$, $p < .001$, $d = 0.97$). There was no significant difference between VF and DI students.

For the instruction phase, there was a significant multivariate effect of condition on mental effort scores ($F[2, 106] = 18.56$, $p < .001$). Planned pairwise comparisons showed that PF students reported significantly greater mental effort than VF students ($F[1, 71] = 12.63$, $p = .001$, $d = 0.84$), as well as DI students ($F[1, 73] = 35.99$, $p < .001$, $d = 1.39$). Vicarious failure students reported significantly greater mental effort than DI students ($F[1, 72] = 6.37$, $p = .014$, $d = 0.58$). There were no significant differences on engagement scores between the conditions during the two phases.

3.2.2. Pre–posttest results

3.2.2.1. Math ability and prior knowledge: ANOVAs with experimental condition as the between-subjects factor revealed no significant difference between the two conditions on math ability ($F[2, 108] = 0.05$, $p = .950$) and prior knowledge ($F[2, 108] = 0.29$, $p = .751$).

3.2.2.2. Posttest: A MANCOVA with scores on procedural knowledge, conceptual understanding, and transfer as the three dependent variables, experimental condition as the between-subjects factor, and math ability, prior knowledge, average engagement score, and average mental effort score as the four covariates revealed significant multivariate main effects of math ability ($F[3, 102] = 6.85$, $p < .001$) and experimental condition ($F[3, 102] = 11.49$, $p < .001$). There were no other significant effects of the covariates or any interaction effects. The model was not sensitive to the exclusion of engagement and mental effort scores.

Univariate ANCOVAs revealed no significant difference between the three conditions on procedural knowledge ($F[2, 104] = 0.25$, $p = .783$). However, there was a significant difference between the three conditions on conceptual understanding ($F[2, 104] = 36.21$, $p < .001$) and transfer ($F[2, 104] = 11.46$, $p < .001$). Planned pairwise comparisons revealed that:

1. Productive failure students significantly outperformed VF students on conceptual understanding ($F[1, 69] = 36.73$, $p < .001$, $d = 1.35$) and transfer ($F[1, 69] = 28.78$, $p < .001$, $d = 1.23$).
2. Productive failure students significantly outperformed DI students on conceptual understanding ($F[1, 71] = 106.47$, $p < .001$, $d = 2.25$) and transfer ($F[1, 71] = 30.62$, $p < .001$, $d = 1.29$).
3. Vicarious failure students significantly outperformed DI students only on conceptual understanding ($F[1, 70] = 12.99$, $p = .001$, $d = 0.80$).

Finally, the number of solutions generated by PF students during their problem-solving phase was significantly correlated with their posttest scores on conceptual understanding ($r[37] = .82$, $p < .001$) and transfer ($r[37] = .88$, $p < .001$), but not with procedural knowledge. There were no such significant correlations in the DI condition. The numbers

of valid or invalid evaluations or their sum did not significantly correlate with the posttest scores.

4. General discussion

Study 2 replicated the findings of Study 1, and together they clearly demonstrate that PF is a more effective teaching method than DI. These findings are consistent with the general math education literature that emphasizes the role of struggle in learning (e.g., Hiebert & Grouws, 2007). Findings are also consistent with and build upon recent work showing that problem solving prior to instruction is more effective than the other way around (DeCaro & Rittle-Johnson, 2012; Schwartz et al., 2011).

Specifically, Studies 1 and 2 make novel contributions in at least three ways. First, Study 2 shows that in addition to preparatory activities such as exploratory problem solving and inventing with contrasting cases, studying and evaluating peer-generated solutions prior to instruction can also be more effective than DI. However, a more significant and novel contribution of Study 2 is that it helps differentiate and demonstrate the greater efficacy of engaging in the cognitive processes of generation and exploration of solutions than studying and evaluating the solutions. Second, both studies evidence a theoretically important link between variability in student production within the problem-solving phase and variability in their learning outcomes. Finally, in contrast to recent work where some form of guidance was provided during the problem-solving phase (e.g., DeCaro and Rittle-Johnson, 2012 provided accuracy feedback, Kapur, 2012 and Schwartz et al., 2011 provided peer collaborative support), Studies 1 and 2 did not provide any such guidance. Given that the proponents of DI have argued that collaborative support or feedback provided during problem solving can help manage the high cognitive load during problem solving, Studies 1 and 2 provide an important demonstration of the effectiveness of unguided problem solving prior to instruction, even though such problem solving may invoke a high cognitive load and not result in correct solutions. Of course, this is not to be taken as an argument that failure during problem solving is a necessary condition for learning. Instead, integrating findings from Studies 1 and 2 with recent research suggests that both unguided and guided problem solving prior to instruction seem to be more effective than DI.

One explanation of the better performance of PF students comes by way of prior knowledge activation and differentiation during the problem-solving phase, which may help them better notice and attend to the critical features of the concept—an explanation that is consistent with those put forth in past research (DeCaro & Rittle-Johnson, 2012; Schwartz et al., 2011; Siegler, 1994). To support this explanation, one must minimally evidence two things: (a) students who were given opportunities to activate and differentiate prior knowledge prior to instruction learned better on average than those who did not. The better performance of PF over DI students on conceptual understanding (items that targeted critical features of the concept) in both the studies provides such evidence. For example, a student having generated a solution of summing the year-on-year deviations

may better understand the need to take deviations from the mean as opposed to some other point in the distribution than a student who did not have the opportunity to generate such a solution, and (b) the greater the prior knowledge activation and differentiation in students who were given such an opportunity, the better the learning on average. To the extent that student-generated solutions and evaluations can be seen as a proxy indicator of prior knowledge activation and differentiation (or as Schwartz and colleagues refer to it as an indicator of a build-up of prior knowledge), the significant relationship between the number of student-generated solutions and learning outcomes seems to provide such evidence.

Given that there was no difference between the three conditions on procedural knowledge (items that targeted the ability to deploy the correct procedures for calculating and interpreting SD), the explanation that PF may have increased the likelihood of correct procedures being selected over incorrect ones was not evidenced. Finally, the explanation by way of better engagement was also not evidenced as there was no significant difference between the conditions on engagement.

It could also be argued that generating solutions may help manage cognitive load because students have to rely on their prior knowledge to do so. According to cognitive load theory (Sweller & Chandler, 1991), if information from the long-term memory (e.g., prior knowledge) can be brought to bear on problem solving, the constraints of working memory can be better managed (Kirschner et al., 2006). Evidence from the two studies, however, did not support this argument. Productive failure students reported significantly greater cognitive load than VF and DI students.

Therefore, there is a need to explain why PF students learned better than VF and DI students in spite of reporting higher cognitive load. A simple explanation could be that there is a trade-off between adverse effects of higher cognitive load and the facilitative effects of prior knowledge activation and differentiation; the latter outweighing the former provided the cognitive load is not so high that learners give up. Thus conceived, high cognitive load may not be monotonically bad for learning (Hiebert & Grouws, 2007; Schmidt & Bjork, 1992). In addition to designing for an optimal amount of cognitive load, a synthesis across several studies suggests other key design features for PF for its benefits to be realized: (a) the problem must admit multiple solutions, strategies, and representations, that is, afford sufficient problem and solution spaces for exploration, (b) the problem should activate learner's prior knowledge—formal as well as intuitive—to solve the problem. Whether or the extent to which the learner is able to correctly solve the problem will depend in part upon the amount and nature of guidance provided, (c) students must themselves generate and explore solutions and not simply be presented with peers' solutions (for a fuller explication of design principles of PF, see Kapur & Bielaczyc, 2012). Over the course of several studies, these features have been used to design and test PF activities in a range of topics in mathematics (e.g., ratio and proportion, SD, normalization) as well as science (e.g., Newtonian kinematics, electric current, genetics).

Interestingly, studies 1 and 2 also showed that students generated more solutions to the problem before receiving instruction than after. This possibly suggests that although

instruction may guide students to produce correct solutions, it may also create a lock-in and constrain search for new solutions. These findings are consistent with the work of Bonawitz et al. (2011), who demonstrated a similar effect on children playing with toys with versus without guidance from adults. An explanation that Bonawitz et al. (2011) proposed is that students may infer from instruction by a knowledgeable adult that all the relevant knowledge and procedures that they need to learn have already been taught during instruction. Such an inference, on one hand, reduces the likelihood of search for irrelevant solutions but, on the other hand, also comes at the expense of limiting exploration.

Finally, it could be argued that the better performance of PF students could be attributed to the recency of instruction they received. There are several reasons why recency may not sufficiently explain the findings. First, performance on the transfer items suggests that the better performance of PF students cannot be due to recency because the content required to solve this item was not covered during instruction. Likewise, the content for five of the eleven conceptual understanding items (items 4a-e identified in the supplementary materials) was not explicitly covered during instruction but had to be deduced from it. Second, similar findings were obtained in prior quasi-experimental comparisons between PF and DI (Kapur, 2010, 2012, 2013), wherein the posttest was not administered immediately after instruction but 1–2 days later. Third, note that DI students engaged in problem solving that tested their ability to apply the concept and procedures of SD (and not on an unrelated task), and findings suggest that they did so successfully. Taken together, these reasons mitigate the possibility that the better performance of PF students was mainly due to recency.

In conclusion, these findings simply suggest that when learning a new concept and its associated procedures, we seem to learn better from our own failed solutions than those of others, although, absent the opportunity to learn from our own failures, we are better off trying to learn from others' failed solutions than from DI.

Acknowledgments

Research reported in this paper was funded by an Office of Educational Research grant from the National Institute of Education (Singapore) to the author. The author would like to thank the school principal, teachers, and students who participated in this study.

Note

1. It must be acknowledged that performance on procedural knowledge was generally at ceiling, and therefore, the posttest may not have been able to pick differences on procedural knowledge. A ceiling effect is in part due to the relatively straightforward nature of computing and interpreting SD.

References

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. H., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*, 322–330.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience and school*. Washington, D.C.: National Academy Press.

DeCaro, M. S., & Rittle-Johnson, B. (2012). Exploring mathematics problems prepares children to learn from instruction. *Journal of Experimental Child Psychology*, *113*(4), 552–568.

Durkin, K., & Rittle-Johnson, B. (2012). The effectiveness of using incorrect examples to support learning about decimal magnitude. *Learning and Instruction*, *22*, 206–214.

Hardiman, P., Pollatsek, A., & Weil, A. (1986). Learning to understand the balance beam. *Cognition and Instruction*, *3*, 1–30.

Hiebert, J., & Grouws, D. A. (2007). The effects of classroom mathematics teaching on students' learning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 371–404). Charlotte, NC: Information Age.

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). Expertise reversal effect. *Educational Psychologist*, *38*, 23–31.

Kapur, M. (2010). Productive failure in mathematical problem solving. *Instructional Science*, *38*(6), 523–550.

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

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

Kapur, M. (2013). Comparing learning from productive failure and vicarious failure . *The Journal of the Learning Sciences*. DOI: 10.1080/10508406.2013.819000

Kapur, M., & Rummel, N. (2012). Productive failure in learning and problem solving. *Instructional Science*, *40*(4), 645–650.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work. *Educational Psychologist*, *41*(2), 75–86.

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science*, *15*(10), 661–667.

Paas, F. (1992). Training strategies for achieving transfer of problem-solving skill in statistics: A cognitive load approach. *Journal of Educational Psychology*, *84*(4), 429–434.

Rittle-Johnson, B., & Star, J. R. (2009). Compared to what? The effects of different comparisons on conceptual knowledge and procedural flexibility for equation solving. *Journal of Educational Psychology*, *101*(3), 529–544.

Roll, I., Aleven, V., & Koedinger, K. (2011). Outcomes and mechanisms of transfer in invention activities. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2824–2829). Austin, TX: Cognitive Science Society.

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), 207–217.

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

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.

diSessa, A. A., Hammer, D., Sherin, B., & Kolpakowski, T. (1991). Inventing graphing: Meta-representational expertise in children. *Journal of Mathematical Behavior*, *10*(2), 117–160.

Siegler, R. S. (1994). Cognitive variability: A key to understanding cognitive development. *Current Directions in Psychological Sciences*, *3*, 1–5.

Siegler, R. S. (2002). Microgenetic studies of self-explanation. In N. Garnott & J. Parziale (Eds.), *Microdevelopment: A process-oriented perspective for studying development and learning* (pp. 31–58). Cambridge, UK: Cambridge University Press.

Sweller, J., & Chandler, P. (1991). Evidence for cognitive load theory. *Cognition and Instruction*, *8*(4), 351–362.

Terwel, J., van Oers, B., van Dijk, I. M. A. W., & van den Eeden, P. (2009). Are representations to be provided or generated in primary mathematics education? Effects on transfer. *Educational Research and Evaluation*, *15*(1), 25–44.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix A: Description of coding of student-generated solutions and evaluations

Appendix B: Additional description of instruction phase

Figure S1: Vicarious failure condition stimuli: Measure A for evaluation training

Figure S2: Vicarious failure condition stimuli: Measures B and C

Figure S3: Vicarious failure condition stimuli: Measures D to F

Table S1: Examples of valid and invalid responses for study and evaluate task