Classroom-based Experiments in Productive Failure

Manu Kapur & Katerine Bielaczyc
National Institute of Education, Singapore

Abstract

We present evidence from three quasi-experimental studies on *productive failure.* In Experiment 1, students experienced either direct instruction (DI) or productive failure (PF), wherein they were first asked to generate a quantitative index for variance before receiving direct instruction on the concept. Experiment 2 examined if it was necessary for students to generate solutions or can these solutions be simply given to the students to study and evaluate. Experiment 3 examined if it was necessary for students to generate solutions before receiving the critical features of the targeted concept, or would simply telling the critical features without any such generation work just as well. In Experiment 1, PF students performed on par with DI students on procedural fluency, and significantly outperformed them on data analysis and conceptual insight items. In Experiment 2, only the effects on conceptual insight and near transfer were significant. In Experiment 3, only the effect on conceptual insight remained significant. Overall, these results challenge the claim that that direct instruction alone is the most effective approach for teaching novel concepts to learners.

Introduction

Proponents of direct instruction bring to bear substantive empirical evidence against un-guided or minimally-guided instruction to claim that there is little efficacy in having learners solve problems that target novel concepts, and that learners should receive direct instruction on the concepts before any problem solving (Kirschner, Sweller, & Clark, 2006). Kirschner et al. (2006) argued that "Controlled experiments almost uniformly indicate that when dealing with novel information, learners should be explicitly shown what to do and how to do it" (p. 79). Based on cognitive load theory, commonly-cited problems with un-guided or minimally-guided instruction include increased working memory load that interferes with schema formation (Tuovinen & Sweller, 1999; Sweller, 1988), encoding of errors and misconceptions (Brown & Campione, 1994), lack of adequate practice and elaboration (Klahr & Nigam, 2004), as well as affective problems of frustration and de-motivation (Hardiman et al., 1986).

Klahr & Nigam's (2004) often-cited study compared the relative effectiveness of discovery learning and direct instruction approaches on learning the control of variable strategy (CVS) in scientific experimentation. On the acquisition of basic CVS skill as well as ability to transfer the skill to evaluate the design of science experiments, their findings suggested that students in the direct instruction condition who were explicitly taught how to design un-confounded experiments outperformed their counterparts in the discovery learning condition who were simply left alone to design experiments without any instructional structure or feedback from the instructor. Further experiments by Klahr and colleagues (e.g., Strand-Cary & Klahr, 2008), and others as well have largely bolstered the ineffectiveness of discovery learning compared with direct instruction (for reviews, see Kirschner et al., 2006).

However, we question whether there is little efficacy in having learners solve problems that target concepts they have not learnt yet. To determine if there such an efficacy, a stricter comparison for direct instruction would be to compare it with an approach where students first generate representations and methods on their own followed by direct instruction. As it can be expected, the generation process will invariably lead to failure, that is, students are rarely able to solve the problems and discover the canonical solutions by themselves. Yet, this very process can be productive for learning *provided* direct instruction on the targeted concepts is subsequently provided (Schwartz & Martin, 2004). As a case in point, we present evidence from our research program on *productive failure* (Kapur, 2008).

Designing for Productive Failure

Productive failure focuses on engaging students in processes that serve two critical cognitive functions, which in turn, prepare students for subsequent direct instruction: a) activating and differentiating prior knowledge in relation to the targeted concepts, and b) affording attention to critical features of the targeted concepts. PF comprises two phases—a generation and exploration phase followed by a direct instruction phase. In the generation and exploration phase, the focus is on affording students the opportunity to leverage their formal as well as intuitive prior knowledge and resources to generate a diversity of solutions for a complex problem; a problem that targets concepts that they have not yet learnt. Research suggests that students do have rich *constructive resources* (diSessa & Sherin, 2000) to generate a variety of solutions for novel problems. At the same time, research also suggests that one cannot expect students, who are novices to the target content, to somehow generate or discover the canonical representations and domain-specific methods for solving the problem (Kirschner et al., 2006).

However, the expectation for the generation and exploration phase is not for students to be able to solve the problem successfully. Instead, it is to generate and explore the affordances and constraints of a diversity of solutions for solving the problem. Our hypothesis is that this process both activates and differentiates prior knowledge (as evidenced in the diversity of student-generated solutions). Furthermore, a comparison and contrast between the various solutions affords opportunities to attend to critical features of the targeted concept. Consequently, the generation and exploration phase provides the necessary foundation for developing deeper understanding of the canonical concept during direct instruction (Kapur, 2009, 2010a/b; Schwartz & Martin, 2004).

Purpose

The purpose of this paper is to report three quasi-experimental studies that help unpack the efficacy of the productive failure (PF) effect. In Experiment 1, we compare PF with direct instruction (DI) to show that PF engenders better prior knowledge differentiation (as evidenced in student-generated solutions), and affords opportunities for students to attend to critical features of the targeted concept. Experiment 2 tests whether prior knowledge differentiation can be engendered by simply giving student-generated solutions to the students to study and evaluate. Finally, Experiment 3 examines the extent to which attention to and understanding of critical features is contingent upon having students go through the generation and exploration phase, or could these critical features simply be told to students as part of direct instruction.

Experiment 1: PF vs. DI

Participants

Participants were 74, ninth-grade mathematics students (14-15 year olds) from two intact classes in an all-boys public school in Singapore. In all three experiments reported in this paper, students were almost all of Chinese ethnicity.

Research Design

A quasi-experimental, pre-post design was used with one class (n = 39) assigned to the 'Productive Failure' (PF) condition, and the other class (n = 35) to the 'Direct Instruction' (DI) condition. Both classes were taught by the same teacher.

Pretest First, all students took a five-item paper and pencil pretest (α = .75) on the concept of variance.

Intervention Next, all classes participated in four, 55-minute periods of instruction on the concept as appropriate to their assigned condition.

In the DI condition, the teacher first explained the concept of variance, and its canonical formulation as the square of the standard deviation ($SD^2 = \sum_1^n \left(x_i - \bar{x}\right)^2 \Big/ n$) using a data analysis problem. Next, the teacher modeled the application of the concept by working through several data analysis problems, highlighting common errors and misconceptions, and drawing attention to critical features of the concept in the process. The data analysis problems required students to compare the variability in 2-3 given data sets, for example, comparing the variability in rainfall in two different months of a year, or comparing the consistency of performance of three soccer players, and so on. To ensure students were engaged and motivated throughout, they were told that they will be asked to solve isomorphic problems after the teacher had worked through the examples with the class. Thereafter, students worked face-to-face in triads on more data analysis problems so that they could benefit from the processes of explanation and elaboration afforded by collaboration. The teacher then discussed the solutions with the class. After each period, students were given isomorphic data analysis problems for homework, which the teacher marked and returned to the students, usually by the following period.

The PF condition differed from the DI condition in one important aspect. Instead of receiving direct instruction upfront, students spent two periods working face-to-face in triads to solve one of the data analysis problems on their own. The data analysis problem presented a distribution of goals scored each year by three soccer players over a twenty-year period. Students were asked to design a quantitative index to determine the most consistent player. During this generation phase, no instructional support or scaffolds were provided. Following this, two periods were spent on direct instruction where the teacher first consolidated by comparing and contrasting student-generated solutions with each other, and then explained the canonical solution just like in the DI condition. Note that because students in the PF condition spent the first two periods generating an index for variance, they solved fewer data analysis problems overall than their counterparts in the DI condition. To make this contrast even sharper, PF students did not receive any homework.

After the second and fourth periods, students from all classes took a five-item, five-point (1=low to 5=high) Likert scale engagement survey (α = .79).

Posttest All students took a five-item, paper and pencil posttest (α = .74) comprising:

i. one item on procedural fluency (calculating SD for a given dataset),
ii. two items on data analysis (comparing means and SDs of two samples; these items were isomorphic with the data analysis problems covered during instruction), and
iii. two items on conceptual insight (required students to evaluate sub-optimal solutions; one item dealing with sensitivity to ordering of data points, and another with outliers)

Maximum score for each of the three types of items was 10; two raters independently scored the items using a rubric with an inter-rater reliability of .92.

Results

Process PF groups generated on average 7 solutions (M = 6.98, SD = 2.48) to the problem. Four categories emerged:

a. *Central tendencies* (e.g., using mean, median, mode);
b. *Qualitative methods* (e.g., organizing data using dot diagrams, frequency polygons, line graphs to examine clustering and fluctuations patterns);
c. *Frequency methods* (e.g., counting the frequency with which a player scored above, below, and at the mean to argue that the greater the frequency at the mean relative to away from the mean, the better the consistency); and
d. *Deviation methods* (e.g., range; calculating the sum of year-on-year deviations to argue that the greater the sum, the lower the consistency; calculating absolute deviations to avoid deviations of opposite signs cancelling each other; calculating the average instead of the sum of the deviations).

Elsewhere, we have described these student-generated solutions in greater detail (Kapur, 2010b). Note that none of the groups were able to generate the canonical formulation of SD. In contrast, analysis of DI students' classroom work revealed that students relied *only* on the canonical formulation to solve data analysis problems. This was not surprising given that had been taught the canonical formulation of SD, which is also easy to compute and apply. All students were accurately able to apply the concept of SD to solve the very problem that the PF students tried to generate a solution to. Finally, on the mean of the two self-reported engagement ratings, there was no difference between the PF condition, M = 3.84, SD = .51, and the DI condition, M = 3.82, SD = .43.

These process findings serve as a manipulation check demonstrating that students in the PF condition experienced "failure," at least in the conventional sense of not being able to generate the canonical solutions. In contrast, DI students were not only just as engaged as PF students but also demonstrated successful application of the canonical formulation to solve data analysis problems, including the one that the PF students solved during the generation phase. The high engagement ratings and performance results also suggest that the DI condition was not simply a case of poor instruction.

Outcome On the pretest, no student demonstrated canonical knowledge of SD, and there was no significant difference between the conditions, $F(1, 72) = 2.56$, $p = .114$.

Posttest performance on the three types of items formed the three dependent variables. Controlling for the effect of prior knowledge as measured by the pretest, $F(4, 134) = 1.89$, $p = .112$, a MANCOVA revealed a significant multivariate effect of condition, $F(4, 134) = 16.802$, $p < .001$, partial $\eta^2 = .33$. Interaction between prior knowledge and experimental condition was not significant.

Table 1: Experiment 1 posttest performance by item type

Experiment 1	PF M *(SD)*	DI M *(SD)*	p / η^2
ProceduralFluency	8.70 (*2.07*)	8.69 (*2.19*)	*ns*
Data Analysis	7.39 (*1.94*)	5.97 (*2.48*)	.013*/.09
Conceptual Insight	6.12 (*2.38*)	3.01 (*1.93*)	.001*/.31

PF students significantly outperformed their DI counterparts on data analysis and conceptual insight problems without compromising on procedural fluency.

Discussion

As hypothesized, the PF design invoked learning processes that not only activated but also differentiated students' prior knowledge (as evidenced by the diversity of student-generated solutions). Whereas PF students worked with the solutions that they generated and the canonical solutions (that they received during direct instruction), DI students worked with only the canonical ones. Hence, DI students worked with a smaller diversity of solutions, and consequently, their prior knowledge was arguably not as differentiated as their PF counterparts. This was a significant difference between the conditions by design. Proponents of DI have repeatedly questioned the utility of getting students to solve novel problems on their own. Instead, they argue that students should be given the canonical solutions (either through worked examples or direct instruction) before getting them to apply these to solve problems on their own (Sweller, 2010).

Experiment 1's findings suggest that there is in fact a utility in having students solve novel problems first. What prior knowledge differentiation affords in part is a comparison and contrast between the various solutions—among the student-generated solutions as well as between the student-generated and canonical solutions. Specifically, these contrasts afford opportunities to attend to the following critical features of the targeted concept that are necessary to develop a deep understanding of the concept:

1. What is the difference between the mean and the distribution around the mean?
2. What is the difference between a qualitative description of the data (e.g., dot diagram, line graphs) and a quantitative description (e.g., range, SD)?
3. What is the difference between the frequency of a point and its position relative to a fixed reference point?
4. Why must we take deviations from a fixed point?
5. Why is the mean usually the fixed point; why can't it be the maximum or the minimum point, or even the median or the mode?
6. Why must we take deviations from the mean for all the points; why not just choose the maximum and the minimum point, or simply the range?
7. Why must deviations from the mean be made positive?
8. Why must we divide the sum of the squared deviations by n; why not simply work with their or sum?
9. Why must we take the square root of the average of the squared deviations?
10. How do outliers affect SD?

However, Experiment 1 raises two further questions:

1. If exposure to both student-generated and canonical solutions is what is essential, then instead of getting students to generate solutions, why not simply let students study the student-generated solutions first (e.g., in the form of well-designed worked examples) and then give them the canonical solutions through direct instruction? Simply put, is it really necessary for students to generate the solutions or can these be given to them? Experiment 2 addresses this question.
2. If what is essential is that students attend to the ten critical features, then why not simply tell students these critical features? Why bother having them generate, and compare and contrast the solutions? Simply put, do students really need to generate before receiving the critical features, or would telling the critical features without any generation work just as well? Experiment 3 addresses this question.

Experiment 2: PF vs. Evaluation

The purpose of Experiment 2 was to examine the difference between: a) having students generate solutions to solve a novel problem, and b) having them study and evaluate student-generated solutions (also see Roll, 2009).

Participants

Participants were 54, ninth-grade mathematics students (14-15 year olds) from two intact classes in an all-boys public school in Singapore.

Research Design

One class (n = 31) was assigned to the PF condition, and the other class (n = 23) to the 'Evaluation' (EV) condition. Both classes were taught by the same teacher. The PF condition was exactly the same as in Experiment 1. The EV condition differed from the PF condition in one important aspect: The generation phase was replaced with an evaluation phase; the subsequent direct instruction phase was the same as in the PF condition.

Whereas PF students had to collaboratively generate solutions to solve the complex problem during the first two periods, EV students took the same two periods to collaboratively study and evaluate the peer-generated solutions (available from Experiment 1). To ensure that students were motivated to understand the given solutions, students were asked to evaluate and rank order the solutions so that they would indirectly be forced to compare and contrast the solutions. 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, seven. The most frequently-generated solutions by the PF students were chosen for EV condition, and none of the chosen solutions contained misconceptions. The seven solutions included one on central tendencies, two on qualitative methods (dot diagram and line graph), two on frequency methods (frequency of the mean and frequency of the mean relative to away from the mean), and two on deviation methods (sum of year-on-year deviation with signs, and average year-on-year deviations without signs).

Because student-generated solutions sometimes lack conceptual clarity in their presentation that may make it difficult for other students to understand and evaluate them, they were converted into well-designed worked examples. EV students received these solutions in the form of worked examples one-by-one (counterbalanced for order), and were given approximately 10-12 minutes for each. The remaining time (approximately 30 minutes) was spent on rank ordering the solutions. Finally, to ensure that EV groups understood the student-generated solutions, the teacher and a research assistant conducted an in-situ check for understanding with the EV groups by asking them to explain their understanding of the solutions. Where students needed help in understanding the solutions, it was readily provided because we did not want students' lack of understanding to adversely affect the fidelity of the EV condition.

Results

Process PF groups produced on average just under 7 solutions (M = 6.78, SD = 2.03) to the problem. As expected, these solutions fell into the four broad categories identified earlier. As in Experiment 1, the mean self-reported engagement ratings were, on average, high, and there was no difference between the PF condition, M = 4.07, SD = .61, and the EV condition, M = 4.12, SD = .53.

Outcome On the pretest, no student demonstrated canonical knowledge of SD, and there was no significant difference between the conditions, $F(1, 63) = 1.16$, $p = .285$.

On the posttest (α = .78), an item on near transfer was added to increase the discriminatory power of the posttest. The near transfer item required students to add data points to a given dataset without changing its mean and SD. Two raters independently scored the items using the same rubric as in Experiment 1 with an inter-rater reliability of .95. Performance on the four types of items formed the four dependent variables. Controlling for the effect of prior knowledge as measured by the pretest, $F(4, 48) = 1.04$, p = .398, a MANCOVA revealed a significant multivariate effect of condition, $F(4, 48) = 3.34$, p = .017, partial η^2 = .22. Interaction between prior knowledge and experimental condition was not significant.

Table 2: Experiment 2 posttest performance by item type

Experiment 2	PF M *(SD)*	EV M *(SD)*	p / η^2
Procedural Fluency	9.60 (*0.98*)	9.43 (*1.73*)	*ns*
Data Analysis	9.83 (*0.90*)	9.34 (*2.28*)	*ns*
Conceptual Insight	4.77 (*1.02*)	3.44 (*1.67*)	.001*/.19
Near Transfer	7.50 (*3.35*)	5.08 (*4.73*)	.039*/.08

PF students significantly outperformed their EV counterparts on conceptual insight and near transfer problems without compromising on procedural fluency and data analysis. Consistent with Roll (2009), exposing students to and having them evaluate student-generated solutions does not seem to be as efficacious as having them generate those solutions before direct instruction.

Experiment 3: PF vs. Strong-DI

The purpose of Experiment 3 was to compare PF condition with a strong DI condition, in which the teacher explicitly explains the 10 critical features.

Participants

Participants were 57, ninth-grade mathematics students (14-15 year olds) from two intact classes in an all-boys public school in Singapore.

Research Design

One class (n = 31) was assigned to the PF condition, and the other class (n = 26) to the 'Strong-DI' condition. Both classes were taught by the same teacher. The PF condition was exactly the same as in Experiment 1. The Strong-DI condition was the same as in Experiment 1 except that the teacher drew attention to the ten critical features during instruction. While explaining each step of formulating and calculating SD, the teacher explained the appropriate critical features relevant for that step. For example, when explaining the concept of "deviation of a point from the mean", the teacher discussed why deviations need to be from a fixed point, why the fixed point should be the mean, and why deviations must be positive. During subsequent problem solving and feedback, the teacher repeatedly reinforced these critical features throughout the lessons.

Results

Process PF groups produced on average just over 7 solutions (M = 7.24, SD = 2.56). These solutions fell into the four broad categories identified earlier. DI students relied *only* on the canonical formulation to solve data analysis problems, and all were accurately able to apply the concept to solve the very problem that the PF students tried to generate solutions to. As in Experiments 1 and 2, the engagement ratings were on average high, and there was no difference between the PF condition, M = 4.15, SD = .44, and the Strong-DI condition, M = 4.22, SD = .32.

Outcome On the pretest, no student demonstrated canonical knowledge of SD, and there was no significant difference between the conditions, $F(1, 55) = .25$, $p = .618$.

The posttest (α = .79) was the same as in Experiment 2. Two raters independently scored the items using the same rubric as in Experiment 2 with an inter-rater reliability of .98. Controlling for the effect of prior knowledge as measured by the pretest, $F(4, 51) = .25$, $p = .907$, a MANCOVA revealed a significant multivariate effect of condition, $F(4, 51) = 2.65$, p =.044, partial η^2 = .17. Interaction between prior knowledge and experimental condition was not significant.

Table 3: Experiment 3 posttest performance by item type

Experiment 3	PF M *(SD)*	Strong-DI M *(SD)*	p / η^2
Procedural Fluency	9.50 (*1.01*)	9.69 (*1.00*)	*ns*
Data Analysis	9.84 (*0.90*)	9.81 (*0.98*)	*ns*
Conceptual Insight	4.44 (*1.24*)	3.55 (*1.13*)	.007*/.13
Near Transfer	7.89 (*2.52*)	7.06 (*2.73*)	*ns*

PF students significantly outperformed their Strong-DI counterparts on conceptual insight without compromising on procedural fluency. Effect on data analysis, which was significant in Experiment 1, was no longer significant in Experiment 3. Effect on near transfer, which was significant in Experiment 2, was no longer significant in Experiment 3. It can be concluded that direct instruction on the critical features appears to be helpful indeed. However, PF students still maintained an edge in terms of conceptual insight. Perhaps one could argue that exposure to sub-optimal solutions in the PF condition can alone explain their better performance on conceptual insight items on the posttest. While this explanation cannot be fully ruled out, Experiment 2 helps mitigate this concern because students in the Evaluation condition were also exposed to the sub-optimal solutions but they still did not perform as well as PF students on conceptual insight.

General Discussion

We reported on three quasi-experimental studies that helped unpack the productive failure (PF) effect. Experiment 1 showed that compared to DI, PF a) engendered better prior knowledge differentiation, and b) afforded opportunities to attend to critical features of the concept of variance, which in turn helped PF students better understand the concept when presented by the teacher during direct instruction subsequently (Schwartz & Martin, 2004). Consequently, PF students performed on par with DI students on procedural fluency, but significantly outperformed them on data analysis and conceptual insight. Although the limitations inherent in quasi-experimental studies with intact classrooms cannot be completely mitigated, note that both the conditions were taught by the same teacher for the same amount of time, exposed students to the same materials (except that DI students were exposed to more data analysis problems), and afforded students the opportunity to benefit from collaborative problem solving.

Experiment 2 further examined prior knowledge differentiation by testing whether it was necessary for students to generate solutions themselves (to engender prior knowledge differentiation), or can these solutions be simply given to the students to study and evaluate. Findings suggested that PF students performed significantly better on conceptual insight and near transfer without compromising on procedural fluency and data analysis.

Because Experiment 1 showed that students do not necessarily attend to or notice deep critical features on their own during direct instruction, Experiment 3 examined whether these features could simply be told to students as part of direct instruction, or if it was more effective for students to generate solutions before receiving these critical features. Findings suggested that although direct instruction on the critical features was effective, having students generate solutions first was still better for developing deep conceptual insight.

In sum, therefore, all three experiments suggested that there is indeed an efficacy in having learners generate and explore representations and methods for solving problems on their own even if they do not formally know the underlying concepts needed to solve the problems, and even if such un-supported problem solving leads to failure initially. By failure, we mean that students were unable to generate the canonical solutions by themselves. Of course, one could argue that PF students were not really failing because they were engaged in processes that were germane for learning (Schmidt & Bjork, 1992). However, when we situate the PF design in the argument made by the proponents of DI, the generation process is invariably seen as failure because the proponents of DI question the utility students generating solutions to novel problems. They argue that students should be given the canonical solutions (either through worked examples or direct instruction) before getting them to apply these to solve problems on their own (Sweller, 2010).

Implications of the above findings pose an interesting dilemma for the limits of working memory (WM) capacity as argued by cognitive load theorists: How is it that students who had not learnt the concept of variance were able to generate multiple representations and solutions to a novel, complex problem targeting that concept in the first place? After all, a complex problem should in and of itself impose a heavy cognitive load on a limited WM capacity, let alone one that targets a novel concept.

To resolve this dilemma, one only need realize that the limits of WM only apply to new or yet-to-be learned information not in the long-term memory (LTM) (Sweller, 2010). However, when dealing with previously stored information in the LTM, these limits tend to be mitigated. Indeed, as Kirschner et al. (2006) argued, "Any instructional theory that ignores the limits of working memory when dealing with novel information or ignores the disappearance of those limits when dealing with familiar information is unlikely to be effective" (p. 77).

If the constraints of WM are contingent upon the novelty of information, and novelty is a function of how what a learner already knows (stored in the LTM) is brought to bear on the new concept being learnt, then it follows that activating relevant prior knowledge in the LTM can help mitigate the constraints of WM. This precisely what PF is designed to do: by designing to activate prior knowledge, PF works to mitigate the WM constraints. This may explain why PF students were able to generate a several solutions to the novel problem. Furthermore, it can be argued that once a particular solution is generated, it forms a resource in the LTM for further generation, that is, generated solutions stored in the LTM can potentially interact with the WM to aid more generation. Finally, these generated structures also become a powerful resource in the LTM that can interact with WM and reduce the cognitive load during subsequent direct instruction, thereby resulting in better learning of conceptual features during direct instruction. Both DI and Strong-DI students did not have these LTM resources that they could leverage to learn better from direct instruction. Thus conceived, one can see why the process of evaluating student-generated solutions can impose a higher WM load than actually generating those very solutions, and consequently interfere with learning.

In sum, if what a learner already knows about a concept is a critical determinant of either limiting or expanding the WM capacity, then does not a commitment to cognitive load theory entail a commitment to understanding whether and to what extent the targeted concept is novel to the learner? However, in our reading, this is rarely taken up by the proponents of DI. Their conception of prior knowledge remains limited to canonical domain-specific knowledge, which in turn, constrains one to work within the limiting aspects of the working memory (e.g., Sweller & Cooper, 1985; Paas, 1992). However, if we allow for the possibility that learners may have some prior knowledge and resources about a concept they have yet to learn, could we not design tasks and activity structures to elicit this knowledge, and by activating and working with these priors in the long-term memory, leverage the expandable aspects of working memory capacity? At the very least, this is a theoretical possibility that the cognitive load theory allows for, and one that should be explored.

Acknowledgements

The research reported in this paper was funded by grants to the authors from the National Institute of Education of Singapore. The authors would like to thank Daniel Schwartz, Nikol Rummel, and Katharina Westermann and for their suggestions on the design of the experiments.

References

Brown, A., & Campione, J. (1994). Guided discovery in a community of learners. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice* (pp. 229–270). Cambridge, MA: MIT Press.

Chi, M. T. H., Glaser, R., & Farr, M. J. (1988). *The nature of expertise*. Hillsdale, NJ: Erlbaum

diSessa, A. A., & Sherin, B. L. (2000). Meta-representation: An introduction. *Journal of Mathematical Behavior*, *19*(4), 385–398.

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

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

Kapur, M. (2009). Productive failure in mathematical problem solving. *Instructional Science*, *38*(6), 523-550.

Kapur, M. (2010b). Productive failure in learning the concept of variance. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 2727-2732). Austin, TX: Cognitive Science Society.

Kapur, M. (2010a). A further study of productive failure in mathematical problem solving: Unpacking the design components. *Instructional Science*. DOI: 10.1007/s11251-010-9144-3

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.

Roll, I. (2009). *Structured invention activities to prepare students for future learning: Means, mechanisms, and cognitive processes.* Thesis, Pittsburgh, PA.

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., & Bransford, J. D. (1998). A time for telling. *Cognition and Instruction, 16*(4), 475-522.

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.

Strand-Cary, M., & Klahr, D. (2008). Developing elementary science skills: Instructional effectiveness and path independence. *Cognitive Development*, *23*(4), 488-511.

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

Tuovinen, J. E., & Sweller, J. (1999). A comparison of cognitive load associated with discovery learning and worked examples. *Journal of Educational Psychology*, *91*, 334–341.