Journal of the Learning Sciences

ISSN: 1050-8406 (Print) 1532-7809 (Online) Journal homepage: http://www.tandfonline.com/loi/hlns20

Designing for Productive Failure

Manu Kapur & Katerine Bielaczyc

THE JOURNAL OF THE LEARNING SCIENCES, 21: 45–83, 2012
Copyright © Taylor & Francis Group, LLC
ISSN: 1050-8406 print / 1532-7809 online
DOI: 10.1080/10508406.2011.591717

Designing for Productive Failure

Manu Kapur and Katerine Bielaczyc
Learning Sciences Laboratory
National Institute of Education, Singapore

In this article, we describe the design principles undergirding *productive failure* (PF; M. Kapur, 2008). We then report findings from an ongoing program of research on PF in mathematical problem solving in 3 Singapore public schools with significantly different mathematical ability profiles, ranging from average to lower ability. In the 1st study, 7th-grade mathematics students from intact classes experienced 1 of 2 conditions: (a) PF, in which students collaboratively solved complex problems on average speed *without* any instructional support or scaffolds up until a teacher-led consolidation; or (b) direct instruction (DI), in which the teacher provided strong instructional support, scaffolding, and feedback. Findings suggested that although PF students generated a diversity of linked representations and methods for solving the complex problems, they were ultimately unsuccessful in their problem-solving efforts. Yet despite seemingly failing in their problem-solving efforts, PF students significantly outperformed DI students on the well-structured and complex problems on the posttest. They also demonstrated greater representation flexibility in solving average speed problems involving graphical representations, a representation that was not targeted during instruction. The 2nd and 3rd studies, conducted in schools with students of significantly lower mathematical ability, largely replicated the findings of the 1st study. Findings and implications of PF for theory, design of learning, and future research are discussed.

When and how to design structure during instructional problem-solving activities is a fundamental theoretical and design issue in education and the learning sciences (Tobias & Duffy, 2010). Instructional structure can be operationalized in a variety of forms, such as structuring of the problem itself, scaffolding, instructional facilitation, provision of tools, content support, expert help, and so on (e.g., Hmelo-Silver, Duncan, & Chinn, 2007; Puntambekar & Hübscher, 2005). Thus

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

conceived, instructional structure is designed to constrain or reduce the degrees of freedom in problem-solving activities (Wood, Bruner, & Ross, 1976), thereby increasing the likelihood of novices achieving performance success. Indeed, a vast body of research supports the efficacy of such an approach. This has led some researchers to argue that instruction should be heavily guided, especially at the start, for without it, learning may not take place (e.g., Kirschner, Sweller, & Clark, 2006). Further support for starting with greater structure during instruction with a gradual reduction (or fading) over time as learners gain expertise comes from several quarters (e.g., Hmelo-Silver et al., 2007; Puntambekar & Hübscher, 2005; Wood et al., 1976).

More often than not, therefore, researchers have tended to focus on different methods for structuring learning and problem-solving activities so as to achieve performance success. In contrast, the role of failure in learning and problem solving, much as it is intuitively compelling, remains largely underdetermined and underresearched by comparison (Clifford, 1984; Schmidt & Bjork, 1992). What is perhaps more problematic is that an emphasis on achieving performance success has led in turn to a commonly held belief that there is little efficacy in novices solving problems without the provision of support structures initially. In contrast, our work is grounded in the belief that engaging novices to try, and even fail, at tasks that are beyond their skills and abilities can, under certain conditions, be productive for developing deeper understandings.

This paper is organized into four sections. We start by reviewing the role of failure in learning and problem solving. Drawing from this review, we articulate the design principles for productive failure (PF) and the theoretical conjectures they embody. We then describe the implementation of the design in Singapore schools used to test, albeit partially, these embodied conjectures (Sandoval, 2004). Specifically, we detail a series of classroom-based experiments comparing a PF design with a direct instruction (DI) design in three schools with students of significantly different mathematical ability, ranging from average to lower ability. We conclude by discussing our findings and possible directions for future research.

THE ROLE OF FAILURE IN LEARNING AND PROBLEM SOLVING

Several scholars and research programs have spoken to the role of failure in learning and problem solving (Clifford, 1984). For example, Schmidt and Bjork (1992) reviewed methods used in the training of motor and verbal skills. They argued that "manipulations that maximize performance during training can be detrimental in the long term" (p. 207), and, conversely, conditions that maximize learning in the long term may not be the ones that maximize performance during the training phase. They introduced the notion of "desirable difficulties" to derive implications for learning designs. Such desirable difficulties include designing variation and/or unpredictability during the training phase, interleaving as opposed to blocking practice on a set of targeted concepts, reducing feedback for learners during the training phase, and using tests as events that afford opportunities to learn.

There is also a growing body of supporting empirical evidence in educational research. For example, research on *impasse-driven learning* (Van Lehn, Siler, Murray, Yamauchi, & Baggett, 2003) in coached problem-solving situations provides strong evidence for the role of failure in learning. Successful learning of a principle (e.g., a concept, a physical law) was associated with events when students reached an impasse during problem solving. Conversely, when students did not reach an impasse, learning was rare despite explicit tutor explanations of the target principle. Instead of providing immediate instructional structure (e.g., in the form of feedback, questions, or explanations) when the learner makes a demonstrable error or is "stuck," Van Lehn et al.'s (2003) findings suggest that it may well be more productive to delay that structure up until the student reaches an impasse—a form of failure—and is subsequently unable to generate an adequate way forward.

Echoing such delaying of instructional structure, Schwartz and Bransford's (1998) work on *preparation for future learning* also demonstrated that when undergraduate students examined similarities and differences among contrasting cases representing a target concept, it prepared them to derive greater benefit from a subsequent lecture or reading on that concept. Further evidence for such preparation for future learning can be found in the *inventing to prepare for learning* research by Schwartz and Martin (2004). In a sequence of design experiments on the teaching of descriptive statistics to intellectually advanced students, Schwartz and Martin demonstrated an existence proof for the hidden efficacy of invention activities when such activities preceded DI (e.g., lectures), despite such activities failing to produce canonical conceptions and solutions during the invention phase.

Kapur's (2008) work on *PF* adds further weight to the role of failure in learning and problem solving. In contrast to a substantive amount of research examining students solving ill-structured problems *with* the provision of various support structures and scaffolds, Kapur examined students solving complex, ill-structured problems *without* the provision of any external support structures. He asked 11th-grade student triads from seven high schools in India to solve either ill- or well-structured physics problems in a synchronous, computer-supported collaborative learning environment. After participating in group problem solving, all students individually solved well-structured problems followed by ill-structured problems. Findings revealed that ill-structured group discussions were significantly more complex and divergent than those of their well-structured counterparts, leading to poor group performance as evidenced by the quality of solutions produced by the groups. However, findings also suggested a hidden efficacy in the complex, divergent interactional process even though it seemingly led to failure;

students from groups that solved ill-structured problems outperformed their counterparts in the well-structured condition in solving the subsequent well- and ill-structured problems individually, suggesting a latent productivity in the failure. Kapur (2008) argued that delaying the structure received by students from the ill-structured groups (who solved ill-structured problems collaboratively followed by well-structured problems individually) helped them discern how to structure an ill-structured problem, thereby facilitating a spontaneous transfer of problem-solving skills. The PF effect in computer-supported collaborative learning settings has since been replicated (Kapur & Kinzer, 2009).

These studies are just a few examples from a growing body of research that emphasizes the need to understand conditions under which delaying structure during instruction can enhance learning (e.g., diSessa, Hammer, Sherin, & Kolpakowski, 1991; Lesh & Doerr, 2003; Slamecka & Graf, 1978). The studies support our argument that there is efficacy in *delaying instructional structure* in order for learners to generate conceptions, representations, and understandings, even though such understandings may not be initially correct. These studies, however, indicate more than simply a delay of instructional structure. They also underscore the presence of desirable difficulties and productive learner activity in solving problems. It is this interest in what is present, that is, the features of productive learner activity (even if it results in "failure"), that forms the core of our work.

Based on the literature and our own studies in PF, we have begun to develop a design theory of what needs to be present in student problem-solving contexts in which instructional structure is delayed. We are interested in testing our theoretical conjectures by investigating their embodiment in the design of problem-solving experiences that, although leading to short-term performance failure, are efficacious in the longer term. We describe these design principles and the theoretical conjectures they embody next.

DESIGNING FOR PF

The literature provides insight into why providing instructional structure too early in the problem-solving process can be problematic. First, students often do not have the necessary prior knowledge differentiation to be able to discern and understand the affordances of domain-specific representations and methods given during DI (e.g., Schwartz & Martin, 2004; for a similar argument applied to perceptual learning, see Garner, 1974; Gibson & Gibson, 1955). Second, when concepts, representations, and methods are presented in a well-assembled, structured manner during DI, students may not understand why those concepts, representations, and methods are assembled in the way that they are (Anderson, 2000; Chi, Glaser, & Farr, 1988; diSessa et al., 1991; Schwartz & Bransford, 1998).

Given these two problems, designing for PF requires engaging students in a learning design that embodies four core interdependent mechanisms: (a) activation and differentiation of prior knowledge in relation to the targeted concepts, (b) attention to critical conceptual features of the targeted concepts, (c) explanation and elaboration of these features, and d) organization and assembly of the critical conceptual features into the targeted concepts.

This resulted in a design comprising two phases: a generation and exploration phase (Phase 1) followed by a consolidation phase (Phase 2). Phase 1 affords opportunity for students to generate and explore the affordances and constraints of multiple representations and solution methods (RSMs). Phase 2 affords opportunity for organizing and assembling the relevant student-generated RSMs into canonical RSMs.

The designs of both phases involved decisions concerning the creation of the activities, the participation structures, and the social surround (see Figure 1). These decisions were guided by the following core design principles to embody the aforementioned mechanisms:

1. Create problem-solving contexts that involve working on complex problems that challenge but do not frustrate, rely on prior mathematical resources, and admit multiple RSMs (mechanisms a and b);
2. Provide opportunities for explanation and elaboration (mechanisms b and c); and
3. Provide opportunities to compare and contrast the affordances and constraints of failed or suboptimal RSMs and the assembly of canonical RSMs (mechanisms b–d).

The ways in which these core principles have been implemented in the designs of the two phases are described next.

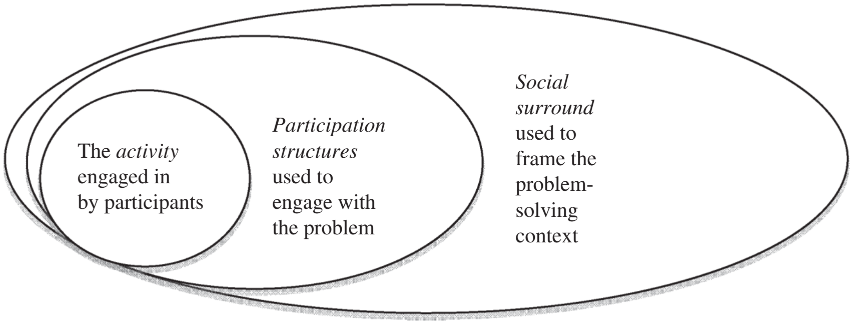

The *activity* engaged in by participants
Participation structures used to engage with the problem
Social surround used to frame the problem-solving context

FIGURE 1 The three layers of the productive failure design.

Phase 1: Generation and Exploration of RSMs

The overall design goal of Phase 1 was to afford opportunities for students to generate and explore a wide variety of RSMs for solving novel, complex problems.

Designing the activity: "sweet-spot" calibration of complex problems. Developing the appropriate problems for PF involves finding a sweet spot where students are challenged yet not frustrated and remain sufficiently engaged in problem solving. Such problem development was approached as a process of calibration, taking into account the complexity of the problems, the mathematical resources of the students, and the affective draw of the way in which a problem is framed.

Complexity of the problems. Well-structured problems commonly found in textbooks typically afford normative RSMs for solving them. In such cases, a learner either is able to solve the problem quickly or simply gives up. In contrast, complex problem scenarios afford multiple RSMs and often require students to make and justify assumptions (Jonassen, 2000; Spiro, Feltovich, Jacobson, & Coulson, 1992; Voss, 1988). Thus, not only do such problem scenarios anchor the learning experience (Brown, Collins, & Duguid, 1989; Cognition and Technology Group at Vanderbilt, 1997), but they also afford opportunities for students to generate a variety of RSMs for solving them (Greeno, Smith, & Moore, 1993). For the unit on average speed, we designed two such complex problem scenarios (see Appendix A for an example).

Prior mathematical resources of students. The complexity of a problem is not the property of the problem alone but a relation between the problem and the problem solver (Lobato, 2003). A problem may well be a simple one for one group of students but not for another. Even more relevant in the present context is that the range of possible RSMs generated by students depends upon the prior mathematical resources students can draw upon. For example, diSessa et al. (1991) found that when sixth graders were asked to invent static representations of motion, students generated and critiqued numerous representations and, in the process, demonstrated not only design and conceptual competence but also meta-representational competence. This suggests that, when given an opportunity, students do have rich *constructive resources* (diSessa & Sherin, 2000) to generate a variety of RSMs for solving problems.

The Grade 7 students in our studies had not had any formal instruction on the targeted concept of average speed. However, in primary school (Grades 1–6), students had been taught concepts such as speed, ratio and proportions, lowest common multiple (LCM), and highest common factor (HCF) as well as problem-solving heuristics such as trial and error (also known as guess and check). Concepts such as speed, ratio and proportion, and domain-general heuristics such as trial and error constituted the set of prior mathematical resources and knowledge necessary for students to be able to attempt complex problems on average speed. Our pilot tests revealed that students were able to use such prior knowledge concepts and heuristics. By taking into consideration students' prior mathematical resources, we tweaked problem parameters such that the problem could not be solved when students used these very resources. For example, the ratios of the walking and biking speeds were deliberately designed to be different so that the problem could not be solved using ratios.

Affective draw of the problem scenario. Through our pilot studies, we found that students were more engaged and interested in the problem when it was presented in the form of a narrative with dialogue. In other contexts, teachers have suggested that a "comic strip" format would have more appeal (e.g., Kapur & Lee, 2009).

In sum, by taking into account the complexity of the problems, the prior mathematical resources of the students, and the affective draw of the way in which a problem is framed, we used pilot tests to developmentally calibrate the complex problems, including the time allocation for group and individual tasks.

Designing the participation structures: enabling collaboration. As much as it is critical for students to generate a variety of RSMs for solving problems, it is equally critical to engage in discourse about their mathematical affordances and constraints. Because collaborative problem solving has been found to be an enabling mechanism that allows students to share, elaborate, critique, explain, and evaluate shared work (Chi et al., 1988; Scardamalia & Bereiter, 2003), small-group collaboration was used as the participation structure during Phase 1. It is important to note that research also suggests that collaborative activities may further enrich the shared representational and solution spaces (diSessa et al., 1991; Schwartz, 1995).

A contextual factor supporting the design component of collaboration is noteworthy. Singapore's mathematics curriculum emphasizes project work; students are exposed to both short- and long-term collaborative projects in the primary (Grades 1–6) and secondary (Grades 7–10) years. Thus, working in groups to solve problems was an activity structure with which students in the three participating schools were largely comfortable. In other contexts and settings, where collaborative problem solving is an activity structure that is novel to both students and teachers, additional structures may need to be designed to support collaboration, because past research suggests that collaboration does not always materialize by simply putting students into groups (Barron, 2003; Dillenbourg, 2002). Hence, in our work, the grouping of students into small groups was not based simply on randomization but on leveraging teachers' understandings of the social dynamics to maximize the likelihood that group members would work well together in their assigned groups (E. G. Cohen, Lotan, Abram, Scarloss, & Schultz, 2002).

Designing the social surround: creating a safe space to explore. The nature of socio-mathematical expectations and norms in the classroom influences the extent to which students actually engage in problem solving (Cobb, 1995; Cobb, Wood, & Yackel, 1993). In spite of strong evidence from previous work and pilot studies that students are able to generate and explore solutions to complex problems, withholding cognitive support runs counter to the normal practice of what teachers and students are used to. At least in the local context of our work, our experience has been that students are used to seeking assistance from their teachers so much so that they do so even before sufficiently trying to solve problems themselves. At the same time, teachers are just as used to providing assistance when it is asked for so much so that often opportunities for students to generate and explore RSMs are missed—opportunities that, as we have argued, are critical for realizing PF.

We worked with the teachers to not provide assistance when asked for but rather to constantly assure students that it was okay not to be able to solve the complex problems as long as they tried various ways of solving them, especially highlighting to them the fact that there were multiple RSMs for the problems. This setting of expectation was important in light of the local context, wherein the usual norm is getting to the correct answer in the most efficient manner given the curricular time constraints. In other contexts and settings in which socio-mathematical norms are more aligned toward exploration, such efforts may not be as deliberate (e.g., Kapur, 2008).

The three layers of designing the activity, the participation structures, and the social surround are meant to act interdependently to maximize the likelihood of students generating and exploring the affordances and constraints of multiple RSMs to solve complex problems. Generation enabled with collaboration facilitates the core mechanisms of activation and differentiation of prior knowledge, as well as attention to and explanation and elaboration of critical conceptual features. Designing the social surround to create appropriate expectations and norms further facilitates the core mechanisms by creating a safe space for generation and exploration. As described, the role of the teacher was not to provide any cognitive or content-related support but mainly to manage the classroom and provide affective support as part of setting the appropriate expectations and norms for problem solving.

Phase 2: Consolidation and Knowledge Assembly

The overall design goal of Phase 2 was to afford opportunities for students to compare and contrast the affordances and constraints of failed or suboptimal RSMs and the assembly of canonical RSMs.

Designing the activity: examining student-generated and canonical RSMs. The central focus of designing the activity was to work with the teacher to engender a whole-class discussion focused on understanding the affordances and constraints of the various RSMs as well as to compare and contrast student-generated RSMs with canonical ones. This activity afforded students the opportunity to attend to and understand the critical conceptual features of the targeted concepts as well as the assembly of these features into the canonical RSMs.

Designing the participation structures: enhancing engagement. The efficacy of a whole-class comparison and contrast of student-generated and canonical RSMs was contingent upon how the teacher facilitated student participation (Nathan & Kim, 2009). Groups were invited to present their work guided by the teacher's questions for clarification and elaboration. The teacher also paraphrased student explanations to explicitly focus attention on the critical conceptual features and, in the process, invite other students to participate in the discussion by questioning, explaining, and elaborating upon one another. For teachers largely and self-admittedly accustomed to a DI mode, these facilitation strategies are not easily developed or adopted. Hence, a professional development program was carried out to develop the teachers' facilitation skills and strategies.

Designing the social surround: creating a safe space to explore. As argued earlier, the establishment of appropriate socio-mathematical expectations and norms in the classroom is critical to ensuring productive participation and discussion. In a DI mode, there is a tendency for the teacher to be the authority and correct students' mistakes, whereas in PF, teachers set the expectations that the discussion of student-generated RSMs was not to assess them as correct or incorrect. Instead, the expectation set was that the process of coming up with RSMs is an important part of mathematical practice (Thomas & Brown, 2007) and that understanding why and under what conditions some RSMs are better than others is important for developing mathematical understanding (diSessa & Sherin, 2000).

EXAMINING THE PF DESIGN IN SINGAPORE SCHOOL CONTEXTS

Having articulated the theoretical conjectures embodied in the PF design, we now describe the design implementation in a series of classroom-based experiments. The experiments were conducted with Grade 7 students at three mainstream, coeducational, public schools in Singapore. The medium of instruction throughout the Singapore school system is English. Students at these schools typically come from middle-class socioeconomic backgrounds.

TABLE 1
Descriptive Statistics for PSLE Performance Across Schools A, B, and C

		PSLE Math Grade[a]		*PSLE Total Score*[b]	
School	*N*	*M*	*SD*	*M*	*SD*
School A	75	1.39	0.52	235.3	3.25
School B	114	2.50	0.62	209.3	5.95
School C	113	3.00	0.88	203.0	5.85

Note. PSLE = Primary School Leaving Examination.

[a] The lower the mean score, the better the math grade; grade A-star is equivalent to 1 point, A to 2 points, B to 3, and so on.

[b] The higher the score, the better the general ability; the PSLE is an aggregate score for all curriculum subjects out of a maximum of 300. More information on the PSLE can be found at www.moe.edu.sg.

The three schools, hereinafter referred to as Schools A, B, and C, were selected based on the academic ability profile of their student intake as evidenced by the Primary School Leaving Examination (PSLE), the sixth-grade national standardized tests used to gain entry into secondary schools (i.e., Grades 7–10) in Singapore. Table 1 presents the descriptive statistics for the PSLE Math grade and PSLE total score for the three schools.

With the PSLE Math grade and total score as the two dependent variables, a multivariate analysis of variance revealed a significant multivariate effect among the three schools, $F(4, 594) = 437.82$, $p < .001$. Students from School A, on average, achieved the highest PSLE score, followed by those from School B and then School C (see Table 1). This effect was statistically significant, $F(2, 299) = 862.48$, $p < .001$, partial $\eta^2 = .85$. As a rule of thumb, partial $\eta^2 = .01$ is considered a small, .06 a medium, and .14 a large effect size (J. Cohen, 1977). Likewise, students from School A, on average, achieved the highest PSLE Math grade, followed by those from School B and then School C. This effect was statistically significant, $F(2, 299) = 116.96$, $p < .001$, partial $\eta^2 = .44$.

The focus of sampling schools based on academic ability needs to be explained further. Past research has largely focused on students with higher ability (e.g., Kapur, 2008; Schwartz & Bransford, 1998), wherein participants have tended to be students who either have qualified for college or are academically advanced or gifted. It is arguably easier to show an existence proof of PF with high-ability students, students who are academically advanced or college bound. What is perhaps more important for both theory and practice is to design and test the tractability of PF with “mainstream” students, students who are not necessarily academically advanced or college bound. By collaborating with schools with students of average (School A) and below average (Schools B and C) academic ability in math, we hoped for a stricter test for PF.

Comparing PF With DI

To bring about change in classroom practice and pedagogy, especially in a system of high-stakes testing such as Singapore, it is important to compare a new instructional design (e.g., PF) with the design most prevalent in practice (e.g., DI). We describe these designs next and articulate our hypotheses for comparing them.

PF. In the generation and exploration phase (Phase 1), student groups (triads) were asked to solve two complex problems (see Appendix A for an example) and given two periods for each of them. After each of the complex problems, students solved extension problems designed as what-if scenarios corresponding to the group complex problems. The study was carried out as part of regular curriculum time. During these six periods, no extra support or scaffolds were provided during the group or individual problem solving in accordance with the design principles, nor was any homework assigned.

In the consolidation phase (Phase 2), the teacher asked the groups to share their RSMs. The goal was to compare and contrast the affordances and constraints of the student-generated RSMs. The teacher then shared the canonical ways (e.g., using algebra) of representing and solving the problems with the class. While doing so, the teacher drew comparisons and contrasts between the canonical and student-generated RSMs and, in the process, explicated the concept of average speed in the context of the problems. Finally, students practiced three well-structured problems on average speed, and the teacher discussed the solutions to these problems.

DI. Students in the DI class were involved in teacher-led lectures guided by the course workbook. The teacher introduced a concept (e.g., average speed) to the class, worked through a few examples, informed them that they would be required to attempt isomorphic problems subsequently, and encouraged students to ask questions. Following this, students solved isomorphic problems for practice. The teacher then discussed the solutions with the class. For homework, students were asked to continue with the workbook problems. Note that the worked-out examples and practice problems were typically well-structured problems with fully specified parameters and canonical RSMs (see Appendix B for examples). The well-structured problems ranged from simple to moderately difficult. This cycle of lecture, practice/homework, and feedback then repeated itself over the course of the same number of periods as in the PF condition. Thus, the amount of instructional time was held constant for the two conditions. Students worked independently most of the time, although some problems were solved collaboratively.

We hypothesized that the PF design would afford students greater opportunities to activate and differentiate their prior knowledge; attend to, explain, and elaborate upon the critical conceptual features of the concept of average speed; and understand the assembly of these features into the canonical RSMs. Consequently, PF students would be able to construct deeper conceptual understanding of the concept of average speed compared to students in a DI design. A deeper conceptual understanding should result in better performance in solving problems on average speed on the posttest, be they standard well-structured problems often found in textbooks or more complex problems (Kapur, 2009). In addition, we also expected that because PF students would have generated and explored a variety of RSMs, they would also demonstrate better *representational flexibility* in solving problems on average speed on the posttest (Ainsworth, Bibby, & Wood, 2002; Lesh, 1999). By *representational flexibility*, we refer to the extent to which students would be able to flexibly adapt their understanding of the concepts of average speed to solve posttest problems that involved a 2-dimensional graphical representation that was not covered during instruction.

We now describe the three classroom-based experiments designed to test the PF hypothesis. The three experiments focused on the differences or variation between the PF and the DI conditions. Following that, we present a mixed-method analysis of variance within the PF condition to better understand the PF effect.

Participants and Design

A quasi-experimental, pre/post design was used in all three schools. Table 2 presents the participants and the research design used in the three schools. Before the unit, all students took a 30-min, 9-item pretest ($\alpha = .72$) as a measure of prior knowledge of the targeted concepts. After the unit, all students took a 35-min, 5-item posttest ($\alpha = .78$).

Differences Among Schools in Research Design and Procedures

The research design and procedures for School A were exactly as described in the previous section. For School B, the research design and procedures were the exactly the *same as in School A* with the following exceptions: Pilot tests with small groups of students prior to the actual study revealed that School B students' frustration thresholds were lower than those in School A. Therefore, the individual extension problems were removed from the design. The time saved was spent on an additional consolidation lesson, given the significantly lower math ability of School B students compared to those from School A.

The research design and procedures in School C were exactly the *same as in School B* with the exception that curricular time allotted for the unit in this school was considerably less than that in Schools A and B. This was a school-level constraint within which we had to work. Consequently, for Phase 1, there was only enough time for two periods of group problem solving followed by two periods of consolidation for Phase 2. The DI condition also lasted four periods

TABLE 2
Summary of Participants and Research Designs in Schools A, B, and C

School	*Participants*	*Design*
School A	75 Grade 7 students (42 M, 33 F) from 2 intact math classes (1 PF class and 1 DI class)	• Teacher familiar with and participated in PF implementation in the previous year • PF ($n = 36$) and DI ($n = 39$) classes taught by the same teacher • *PF condition:* 6 periods of generation followed by 1 period of consolidation; *DI condition:* 7 periods
School B	114 Grade 7 students (63 M, 51 F) from 3 intact math classes (2 PF classes [PF-A and PF-B] and 1 DI class)	• 2 teachers, both new to implementation • One PF class (PF-A; $n = 38$) and the DI class ($n = 40$) taught by the same teacher; the other PF class (PF-B; $n = 36$) taught by the second teacher • *PF condition:* 4 periods of generation followed by 2 periods of consolidation; *DI condition:* 6 periods
School C	113 Grade 7 students (54 M, 59 F) from 3 intact math classes (2 PF classes [PF-A and PF-B] and 1 DI class)	• 2 teachers, both new to implementation • One PF class (PF-A; $n = 38$) and the DI class ($n = 38$) taught by the same teacher; the other PF class (PF-B; $n = 37$) taught by the second teacher • *PF condition:* 2 periods of generation followed by 2 periods of consolidation; *DI condition:* 4 periods

Note. M = male; F = female; PF = productive failure; DI = direct instruction.

in total—the time allocated for this unit in the curriculum. It is interesting that this gave us an opportunity to test the PF hypothesis for a considerably shortened design intervention, though, unlike in Schools A and B, we were not sure whether our hypotheses would hold in this school.

Data Sources and Analytical Procedures

Data sources and analytical procedures were the same for all three participating schools. Both process and outcome measures were analyzed.

Process measures for the PF condition. Each PF group was given blank sheets of A4 paper for its group work. All group discussions were captured in audio and transcribed by a research assistant. Process measures included the following.

Group/individual performance. The group work artifacts were examined to determine the number of PF groups that were able to solve to the complex problems successfully. There was a clear bimodal distribution (i.e., groups were either able to find a correct solution or not despite their extensive exploration of the problem and solution spaces). Thus, if and only if a group was able to find a correct solution (e.g., the partition distance for the complex problem scenario in Appendix A), it was deemed successful in its problem-solving efforts. The average of the percentages of groups that solved the first problem successfully and those that solved the second problem successfully was taken as the measure of group performance. Note that because students in School A also solved individual extension problems, individual performance was operationalized as the average of the percentages of students who solved the first problem successfully and those who solved the second problem successfully. Students in Schools B and C did not solve extension problems.

Group RSM diversity. The group work artifacts and the discussion transcripts were used to determine the maximal set of RSMs generated by the PF groups. The set of RSMs identified in the group work artifacts was used to chunk the group discussion into smaller episodes. For example, if the group work artifacts revealed that the group used ratios to solve the problem, then the relevant episode from the discussion in which the group discussed the ratios method was identified. Chunking of a discussion into episodes was simplified by the fact that there were generally clear transitions in the discussions when a group moved from one RSM (e.g., ratios, trial and error) to another (e.g., algebra). Episodes containing additional RSMs not captured in the group work artifacts were also identified. These included qualitative insights that were important conceptually for solving the problem. In accordance with the hypothesis, the analysis was focused squarely on RSMs, and episodes of non-task behavior and social talk were not included in the analysis. This process was repeated for all PF groups. Two raters independently chunked the group transcripts into episodes and coded the episodes into RSM type. The interrater reliabilities (Krippendorff's alphas) for chunking of transcripts into episodes and coding of the episodes were .94 and .97, respectively.

A total of *nine* different RSMs emerged from this analysis. *RSM diversity* was defined as the number of different RSMs generated by a group. Thus, a group could score 0 through 9 for RSM diversity; the higher the score, the greater the RSM diversity. The nine RSMs were as follows:

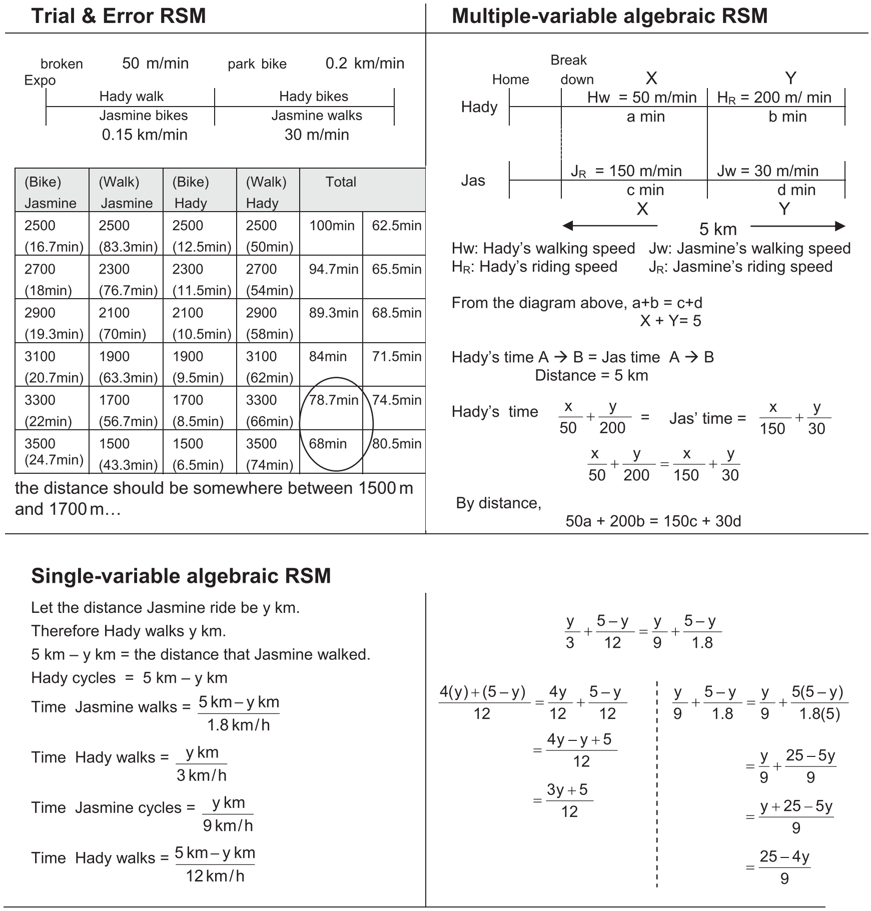
1. *Hady should walk more.* All groups were able to develop the idea that because Hady's biking speed was greater than Jasmine's, he should do more of the walking. This was a qualitative concept that emerged in the group discussions.
2. *Jasmine's walking distance must equal Hady's biking distance and vice versa.* All groups went further to generate the insight that the partitioning of the total distance into walking and biking components should be such that Jasmine's walking distance must equal Hady's biking distance and that Jasmine's biking distance must equal Hady's walking distance.
3. *Diagrams.* All groups were able to draw an accurate diagram to represent the journeys of Jasmine and Hady. These diagrams contained information about distances, speeds, and partition point. For all groups, the diagrams seemed to anchor their problem-solving efforts. Examples of such diagrammatic representations can be found in Figures 2, 3, and 4.
4. *LCM/HCF.* Some groups used their prior knowledge of LCM and HCF to represent and solve the problem. For example, Figure 2 shows how one group attempted to use LCM to solve the problem. The group took the LCM of the biking speeds to determine the shortest distance into which the biking speeds would factor, including the time it would take to do so. However, the group did not pursue this method any further, in part because the LCM did not form a proper factor of the distance to be traveled (i.e., 600 does not divide 5,000 completely). The same group then tried to find the HCF of Jasmine's biking speed and Hady's walking speed. Again, as shown in Figure 2, the strategy was to find the number of times the HCF divided the remaining distance and then apportion the parts accordingly. This method too did not lead to a successful solution.
5. *Ratios.* The use of ratios or proportions was fairly common (see Figure 4). The idea here was simple: Divide the total distance into an appropriate number of parts using the sum of the numerator and the denominator of the

Lowest Common Multiple (LCM) RSM	Highest Common Factor (HCF) RSM
Jasmine's biking speed = 0.15 km/ min (150 m/min) Hady's biking speed = 0.2 km/ min (200 m/min) 8750 m – 3750 m = 5000 m *(remaining distance left)* 2 150 200 LCM 5 75 100 = 2x5x5x4x3 5 15 20 = 600 m 4 3 4 3 3 1 1 1 Time required by Jasmine to bike 600 m = 600 m ÷ 150 m/min = 4 mins	Distance Jasmine bikes for the first 5 minutes = 5 x 150 m/ min = 750m Distance Hady walks for the first 5 minutes = 5 x 50 m/min = 250m HCF of Jasmine's and Hady's distances for the first 5 minutes 10 750 250 HCF 5 75 25 = 10x5x5 5 15 5 = 250 3 1 (or 0.25 km) So, the total remaining distance (5 km) consists of 20 units of 0.25 km. Half of that distance comprises 10 units of 0.25 km. Since Jasmine travels 500 m more than Hady in the first part of the journey, she should be given = 10 + (500 ÷ 250) = 12 units of 0.25 km, or (12 X 0.25) km = 3 km. Time Jasmine takes to bike 3 km = 3000 ÷ 150 m/min = 20 mins Time Hady takes to walk 3 km = 3000 ÷ 50 m/min = 60 mins Difference between both of their times = 60–20 mins = 40 mins

FIGURE 2 Examples of the LCM and HCF methods. RSM = representation and solution method.

ratio, and then apportion the distances in inverse proportion to the speeds. Conceptually speaking, as argued earlier, the method is a reasonable one except that it did not work in the present case because the ratios of the walking and riding speeds were designed to be different.

6. *Trial and Error A (brute force).* The use of trial and error was also fairly common, and the few groups that managed to solve the problem successfully all relied on trial and error. Two versions of trial and error emerged. The first one was what we refer to as the *brute force method*, shown in Figure 3. That is, groups would make an initial guess at the partition point, be it distance or time, typically the midpoint or the starting point, and then increment it systematically until they converged upon a solution.
7. *Trial and Error B (connected to the ratios methods).* The second trial and error method was more sophisticated because it used information from the ratios method. Specifically, the choice of the starting guess for the partition point was informed by the partition distance derived from the ratio method. Therefore, instead of starting from 2.5 km as the initial guess as was the case in the brute force method, groups used the answer from the ratio method (either 2.8 km or 3.125 km) as the initial guess for the trial and error method. This reduced the computational and search load significantly and was consequently a faster method than the brute force method.



Trial & Error RSM

(Bike) Jasmine	(Walk) Jasmine	(Bike) Hady	(Walk) Hady	Total	
2500 (16.7min)	2500 (83.3min)	2500 (12.5min)	2500 (50min)	100min	62.5min
2700 (18min)	2300 (76.7min)	2300 (11.5min)	2700 (54min)	94.7min	65.5min
2900 (19.3min)	2100 (70min)	2100 (10.5min)	2900 (58min)	89.3min	68.5min
3100 (20.7min)	1900 (63.3min)	1900 (9.5min)	3100 (62min)	84min	71.5min
3300 (22min)	1700 (56.7min)	1700 (8.5min)	3300 (66min)	78.7min	74.5min
3500 (24.7min)	1500 (43.3min)	1500 (6.5min)	3500 (74min)	68min	80.5min

the distance should be somewhere between 1500 m and 1700 m…

Multiple-variable algebraic RSM

Hw: Hady's walking speed Jw: Jasmine's walking speed
H_R: Hady's riding speed J_R: Jasmine's riding speed

From the diagram above, a+b = c+d
X + Y= 5

Hady's time A → B = Jas time A → B
Distance = 5 km

Hady's time $\frac{x}{50}+\frac{y}{200}$ = Jas' time = $\frac{x}{150}+\frac{y}{30}$

$$\frac{x}{50}+\frac{y}{200}=\frac{x}{150}+\frac{y}{30}$$

By distance,

$$50a + 200b = 150c + 30d$$

Single-variable algebraic RSM

Let the distance Jasmine ride be y km.
Therefore Hady walks y km.
5 km – y km = the distance that Jasmine walked.
Hady cycles = 5 km – y km

Time Jasmine walks = $\frac{5\,\text{km}-y\,\text{km}}{1.8\,\text{km/h}}$

Time Hady walks = $\frac{y\,\text{km}}{3\,\text{km/h}}$

Time Jasmine cycles = $\frac{y\,\text{km}}{9\,\text{km/h}}$

Time Hady walks = $\frac{5\,\text{km}-y\,\text{km}}{12\,\text{km/h}}$

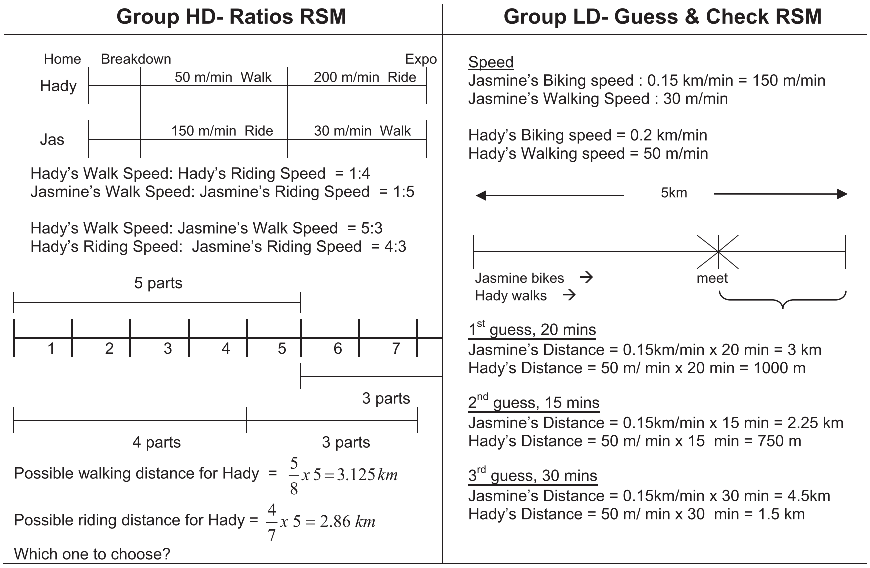
$$\frac{y}{3}+\frac{5-y}{12}=\frac{y}{9}+\frac{5-y}{1.8}$$

$$\frac{4(y)+(5-y)}{12}=\frac{4y}{12}+\frac{5-y}{12}$$
$$=\frac{4y-y+5}{12}$$
$$=\frac{3y+5}{12}$$

$$\frac{y}{9}+\frac{5-y}{1.8}=\frac{y}{9}+\frac{5(5-y)}{1.8(5)}$$
$$=\frac{y}{9}+\frac{25-5y}{9}$$
$$=\frac{y+25-5y}{9}$$
$$=\frac{25-4y}{9}$$

FIGURE 3 Examples of the trial and error, single-variable, and multiple-variable algebraic methods. RSM = representation and solution method.

8. *Letter-Symbolic Algebra A (multiple variables).* As shown in Figure 3, a multiple-variable algebraic representation was one of the two types of algebraic representations that the groups developed. However, there were more variables than equations, which made the system of equations unsolvable.
9. *Letter-Symbolic Algebra B (single variable).* A small proportion of groups was able to derive a single-variable algebraic representation of the problem as shown in Figure 3. However, a lack of algebraic manipulation skills prevented them from being able to solve the equation successfully.



Hady's Walk Speed: Hady's Riding Speed = 1:4
Jasmine's Walk Speed: Jasmine's Riding Speed = 1:5

Hady's Walk Speed: Jasmine's Walk Speed = 5:3
Hady's Riding Speed: Jasmine's Riding Speed = 4:3

Possible walking distance for Hady = $\frac{5}{8} x\, 5 = 3.125\, km$

Possible riding distance for Hady = $\frac{4}{7} x\ 5 = 2.86\ km$

Which one to choose?

Speed
Jasmine's Biking speed : 0.15 km/min = 150 m/min
Jasmine's Walking Speed : 30 m/min

Hady's Biking speed = 0.2 km/min
Hady's Walking speed = 50 m/min

1st guess, 20 mins
Jasmine's Distance = 0.15km/min x 20 min = 3 km
Hady's Distance = 50 m/ min x 20 min = 1000 m

2nd guess, 15 mins
Jasmine's Distance = 0.15km/min x 15 min = 2.25 km
Hady's Distance = 50 m/ min x 15 min = 750 m

3rd guess, 30 mins
Jasmine's Distance = 0.15km/min x 30 min = 4.5km
Hady's Distance = 50 m/ min x 30 min = 1.5 km

FIGURE 4 Group HD's ratio method and Group LD's guess and check method. HD = high diversity; LD = low diversity; RSM = representation and solution method.

Process measures for the DI condition. Performance from the daily homework assignment provided a proxy measure for student performance in the DI condition. The homework comprised six to eight well-structured problems (similar to the problems in Appendix B) that the teacher scored either 1 (if answered correctly) or 0 (if answered incorrectly). Computational or calculation errors were not penalized given our focus on conceptual understanding. The average percentage score across the homework assignments was taken as a measure of DI student performance.

Outcome measures for the PF and DI conditions. The 5-item posttest comprised three well-structured problem items similar to those on the pretest, one complex problem item, and one graphical representation item (see Appendix C for an example of each). The interrater reliability (Krippendorff's alpha) for scoring the posttest was .87. The three types of items formed the three dependent variables in a multivariate analysis of covariance (MANCOVA), with pretest score as the covariate.

We also held debriefing sessions with the teachers after each of the two phases of the PF design. These debriefing sessions were captured in audio and transcribed. Data from these sessions are used only as corroborating evidence to support the discussion of our findings.

RESULTS

Pretest

For School A, there was no significant difference between the PF and DI classes on the pretest, $F(1, 73) = 0.18$, $p = .675$. The same was true for School B, $F(2, 114) = 0.54$, $p = .586$; and School C, $F(2, 110) = 2.34$, $p = .101$.

Process

Table 3 summarizes the findings from the process analysis. With regard to RSM diversity, findings suggest that PF groups in all three schools were able to generate multiple RSMs for solving complex problems. An analysis of covariance revealed a significant difference among schools on RSM diversity, $F(2, 181) = 3.51$, $p = .032$, partial $\eta^2 = .04$. Levene's test was not significant ($p = .467$). Notably, this difference among schools had a small effect size, which is in and of itself a significant finding in light of the large effect size difference in academic ability of the students from the three schools.

With regard to group and individual performance on the complex problems, findings suggest that in spite of generating multiple RSMs, students were ultimately unable to solve the problems successfully either in groups or individually. As can be seen from Table 3, the percentage of groups that managed to solve the

TABLE 3
Descriptive Statistics for Process Measures in the PF and Direct Instruction Conditions

	PF				*Direct Instruction*	
			RSM Diversity		*Individual Homework Performance*[a]	
School	*No. of PF Groups*	*Group/Individual Performance*	*M*	*SD*	*M*	*SD*
School A	12	16%[b]/11.5%[c]	6.83	1.44	91.4%	4.5%
School B	25	7%/NA	6.11	1.36	92.6%	4.3%
School C	26	0%/NA	5.43	1.49	91.5%	5.1%

Note. PF = productive failure; RSM = representation and solution method; NA = not applicable.

[a] Individual homework performance was averaged across four assignments in School A, three in School B, and two in School C. The relatively small size of the standard deviations indicates a generally high level of homework performance.

[b] 16% means that on average about 2 out of the 12 groups were successful in solving the complex problems. All successful solutions used the trial and error method.

[c] Individual extension problems were given in School A only.

complex problems was very low in all three schools. Likewise, individual performance on the extension problems in School A was also poor. In contrast, average individual homework performance in the DI condition was very high in all three schools.

These process findings double up as a manipulation check demonstrating that students in the PF condition experienced failure at least in the conventional sense of performance success and efficiency. In contrast, students in the DI condition, by design, repeatedly experienced performance success in solving well-structured problems under the teacher's close monitoring, scaffolding, and feedback.

Posttest

Table 4 presents the descriptive statistics for the adjusted posttest scores for the PF and DI classes for Schools A, B, and C. The interaction between prior knowledge (covariate) and experimental condition (PF vs. DI) was not significant in School A, $F(3, 69) = 0.44$, $p = .725$; or in School B, $F(3, 108) = 1.39$, $p = .250$; or in School C, $F(3, 107) = 0.21$, $p = .886$. Box's M test for homogeneity of variance was also not significant in any of the three schools. This means that the assumptions of parallelism of regression planes and sphericity were not violated, allowing us in

TABLE 4
Summary of Posttest Performance

		Well-structured Items (Max Score = 19)		*Complex Item (Max Score = 7)*		*Representational Flexibility (Max Score = 3)*	
Variable	*N*	*M*	*SD*	*M*	*SD*	*M*	*SD*
School A	75						
PF	36	16.11	1.91	4.18	2.21	2.70	0.60
DI	39	14.31	2.39	2.97	2.26	2.18	0.84
School B[a]	114						
PF	74	15.76	2.42	3.20	2.56	2.30	0.95
PF-B	36	16.30	2.29	3.86	2.44	2.33	0.93
PF-A	38	15.30	2.45	2.58	2.65	2.26	0.98
DI	40	14.30	3.11	1.01	2.32	0.98	0.97
School C[a]	113						
PF	75	14.09	4.16	1.59	1.84	2.08	0.96
PF-B	37	14.50	4.25	1.90	2.16	2.26	0.95
PF-A	38	13.60	4.08	1.28	1.40	1.89	0.97
DI	38	13.30	4.08	0.68	1.18	1.49	1.00

Note. PF = productive failure; DI = direct instruction.

[a] Two teachers were involved in the implementation. One taught the PF-A and DI classes; the other taught the PF-B class.

TABLE 5
Summary of Multivariate and Univariate Effects of Experimental Condition (Productive Failure vs. Direct Instruction)

Variable	*F*	*p*	*Partial η^2*
School A: Omnibus	$F(3, 70) = 8.57$	<.001*	.27
Well-structured items	$F(1, 72) = 13.54$	<.001*	.16
Complex item	$F(1, 72) = 5.48$	.022*	.07
Representational flexibility item	$F(1, 72) = 10.33$	.002*	.13
School B: Omnibus	$F(3, 109) = 7.82$	<.001*	.18
Well-structured items	$F(1, 111) = 7.06$	.009*	.06
Complex item	$F(1, 111) = 5.89$	.017*	.05
Representational flexibility item	$F(1, 111) = 9.43$	.003*	.08
School C: Omnibus	$F(3, 108) = 3.67$	.014*	.09
Well-structured items	$F(1, 110) < .001$	.987	.00
Complex item	$F(1, 110) = 5.15$	.025*	.05
Representational flexibility item	$F(1, 110) = 5.27$	.024*	.05

*Significant result.

turn to interpret the main effects of prior knowledge and experimental condition in the MANCOVA (Stevens, 2002).

There was a significant multivariate effect of prior knowledge on posttest scores in School A, $F(3, 70) = 4.35, p = .007$, partial $\eta^2 = .16$; and also in School C, $F(3, 108) = 5.98, p = .001$, partial $\eta^2 = .14$. However, in School B there was no significant multivariate effect of prior knowledge on posttest scores, $F(3, 70) = 1.54, p = .207$, partial $\eta^2 = .04$.

Table 5 presents the multivariate (omnibus) and univariate main effects of experimental condition. The multivariate main effect of experimental condition was significant in all three schools. All but one univariate main effect of experimental condition was not significant: In School C, the difference between the PF and DI classes on well-structured items did not reach significance.

Bonferroni-corrected post hoc tests were carried out to examine whether the PF versus DI effect in Schools B and C was mainly due to one or both of the PF classes (see Table 6). On the well-structured items, the only PF classes that significantly outperformed their respective DI classes were the PF class from School A and the PF-B class from School B. On the complex item, all PF classes in the three schools significantly outperformed their respective DI classes. On the graphical representation item, all PF classes except the PF-A class in School C significantly outperformed their respective DI classes. These results suggest that PF was pedagogically tractable in classrooms with significantly lower math and general academic ability. However, differences among schools are difficult to attribute to a particular factor because of multiple confounds between school, teacher, student ability, and intervention specifications.

TABLE 6
Summary of Post Hoc Effects for Schools B and C

	Well-Structured Items		*Complex Item*		*Representational Flexibility*	
Variable	*F*	*p*	*F*	*p*	*F*	*p*
School B						
DI vs. PF-B	8.87	.004*	21.36	.001*	8.54	.005*
DI vs. PF-A	*ns*	*ns*	7.48	.008*	7.94	.006*
School C						
DI vs. PF-B	*ns*	*ns*	7.41	.008*	6.99	.010*
DI vs. PF-A	*ns*	*ns*	4.01	.049*	*ns*	*ns*

Note. DI = direct instruction; PF = productive failure; *ns* = nonsignificant.
*Significant result.

Thus far, our analysis has been aimed at understanding the variance between the PF and DI conditions. This analysis also suggested a variance within the PF condition (especially in Schools B and C), which forms the next focus of our investigation.

VARIATION WITHIN THE PF CONDITION

The aim of unpacking variance within the PF condition is to build explanatory support for the theoretical conjectures embodied in the PF design. The locus of this variation could be in either phase of the design: (a) generation and exploration of RSMs, or (b) consolidation and knowledge assembly, or a combination of both. Although we believe it is critical to unpack variation across the entire design in the long run, for the purposes of this paper and because of constraints on what can be accomplished in a single paper, we focus on unpacking variation in the generation and exploration phase. First we examine whether the diversity of RSMs generated by each group relates to the subsequent posttest performance by members of that group. Based on this relationship, we begin to unpack the actual interactions occurring among group members in generating such RSMs.

We had hypothesized that the extent to which the PF design activates and differentiates students' prior knowledge (as evidenced by the generation of multiple RSMs) will influence the extent to which they learn. Therefore, an examination of the relationship between group RSM diversity and learning outcomes as measured on the posttest across *all* PF groups (63 PF groups: 12 from School A, 25 from School B, and 26 from School C) in the three schools would provide some evidence in support of the hypothesis.

Table 3 presents the descriptive statistics for RSM diversity in the PF groups. After we accounted for variation across schools, $F(6, 354) = 3.79$, $p = .001$, partial $\eta^2 = .06$; and controlled for variation in pretest, $F(3, 178) = 5.83, p < .001$, partial $\eta^2 = .09$; a MANCOVA[1] revealed that RSM diversity had a significant multivariate effect, $F(3, 178) = 230.48$, $p < .001$, partial $\eta^2 = .80$, on the well-structured, $F(1, 180) = 222.15, p < .001$, partial $\eta^2 = .55$; complex, $F(1, 180) = 56.80$, $p < .001$, partial $\eta^2 = .24$; and graphical representation, $F(1, 180) = 6.99$, $p = .009$, partial $\eta^2 = .04$, items on the posttest.

In fact, comparison of *F* values indicates that the order of effect was the greatest for RSM diversity, followed by the pretest and then the school. More specifically, a comparison of the effect sizes suggests that the effect of RSM diversity was about 9 times stronger than the pretest and 13 times stronger than the school.

Contrasting-Case Analysis

The preceding analysis underscores the significant role played by RSM diversity; the greater the RSM diversity, the better, on average, the performance of the group's members. In addition to generating multiple RSMs (mechanism a), the PF design principles emphasize the role of a collaborative activity structure in enabling an exploration and elaboration of these RSMs, in turn affording greater opportunities for attending to, explaining, and elaborating upon the critical features of the targeted concept embedded in the problem (mechanisms b and c). The purpose of the following contrasting-case analysis is to use discussion excerpts from two groups with contrasting levels of RSM diversity and *illustrate* how these groups additionally differed in their exploration of the RSMs they generated and how this difference in exploration potentially influenced opportunities to attend to critical features of the problem.

Selection of contrasting-case groups. Two groups, one with high diversity (hereinafter referred to as Group HD) and another with low diversity (hereinafter referred to as Group LD) from School A that contrasted in their RSM diversity, were selected. Group HD generated six conceptual structures (two qualitative insights, diagrammatic representation, ratios method, guess and check method, and algebraic method), whereas Group LD generated only three (diagrammatic representation, guess and check method, and LCM combined with guess and check method). Consistent with the quantitative analysis in the preceding section, the three members of Group HD (hereinafter referred to as HD1, HD2, and HD3) scored 22, 23, and 25, respectively, on the posttest (maximum

[1] Note that the MANCOVA was carried out treating each individual student as an independent observation. Because students worked in groups, this assumption of independence is not valid, and a MANCOVA may result in a more liberal significance level. Ideally, we would have carried out a multilevel analysis, but the sample size was too small vis-à-vis the number of variables being analyzed. Although this remains a limitation, it is somewhat mitigated by the large *F* values of the multivariate and univariate effects (Hox, 1995).

score = 29), whereas the three members of Group LD (hereinafter referred to as LD1, LD2, and LD3) scored 18, 19, and 19, respectively. Therefore, the contrast on RSM diversity was also a contrast in posttest performance.

In the following contrasting-case analysis, we illustrate how the two groups contrasted in terms of their efforts in exploring one of their RSMs. Our strategy is to present the contrasting excerpts[2] from the two groups followed by an analysis of the contrast. Each excerpt is also accompanied with interpretive comments for each utterance, the mechanisms (a, b, or c) invoked, and the collaboration moves (e.g., proposal, question, evaluation, explanation) made.

Exploring an RSM. Exploration of an RSM requires that groups be able to understand the affordances and constraints of the RSM. This would include (a) deploying and working with appropriate representational forms; (b) carrying out appropriate manipulations and computations; (c) understanding whether the method works and, if it does not, then why it does not; and (d) working together so that all group members can develop a shared understanding of the method. In doing so, a good exploration would help draw attention to critical features of the targeted concept.

The two excerpts from Groups HD and LD (see Tables 7 and 8, respectively) contrast the nature and extent to which each group was able to explore an RSM they generated. Group HD worked on using ratios to solve the problem, whereas Group LD worked on trial and error, as shown in Figure 4. For the purposes of the contrast, the difference between the RSMs is not as important as how the two groups explored them. If anything, the trial and error method guarantees a solution, but the problem was designed in such a way that the ratios method would not result in a solution.

The starting point of the excerpts was determined because the two excerpts immediately followed the groups' attempts to understand the problem and started by proposing a solution. The ending of the excerpt was determined as the utterance that indicated that the group had either reached an impasse or moved on to another method.

Comparing and contrasting the two excerpts reveals that both groups were able to deploy appropriate representational forms and carry out the necessary computations for their respective solution method (also see Figure 4). However, the two groups seemed to be quite different in terms of the mechanisms invoked, which influenced their understanding of the affordances and constraints of the method

[2] Note that the excerpts have undergone some minimal editing for language and grammar to make them more comprehensible. Minimal language and grammar editing was necessary to make them readily accessible to a wide audience because students often used a local variant of English called *Singlish* (short for Singapore English) in their interactions with one another.

TABLE 7
Solution Exploration by the HD Group

Line	*Speaker*	*Utterance*	*Interpretive Comment*	*Mechanism*[a]	*Collaboration*
1	HD1	I know . . . if you draw this line for Hady and this one for Jasmine, then say here the bicycle breaks down, and this is the remaining distance 5 km right? So, Hady has to walk the same distance as Jasmine rides. Then Jasmine walks the same distance as Hady rides. Right?	HD1 uses a diagram to draw attention to a critical feature—the relation between walking and biking distances	a, b, c	Proposal
2	HD3	Yes, but their speeds are different . . .	HD3 agrees and draws attention to another parameter: difference in the speeds	b	Agreement, evaluation
3	HD2	maybe we can break the distance using the LCM?	HD2 proposes a solution method and draws attention to the parameter of total distance	a, b	Proposal
4	HD1	Oh, yes, we can use the ratios, right?	HD1 agrees with HD3 and proposes an alternative solution	a	Agreement, proposal
5	HD2	Yes, that's what I meant	HD2 agrees and clarifies that by *LCM* he meant the ratios method		Clarification
6	HD3	Wait, let me see . . . let's try first	HD3 evaluates and wants to try the proposed solution	c	Evaluation
7	HD2	What do you mean?	HD2 questions, seeks explanation		Question
8	HD3	. . . No, I don't think this will work.	HD3 evaluates that the ratios method will not work	c	Evaluation
9	HD1	Why not?	HD1 seeks an explanation		Question
10	HD3	We can't. If we use the biking speed right, ratio is 4 is to 3, but if we use walking speed, this is 5 is to 3.	HD3 explains and draws attention to a critical features of the ratios—the walking and biking speeds are in different ratios	b, c	Evaluation, explanation
11	HD1	Huh? I don't understand	HD1 still does not understand		Question

(*Continued*)

TABLE 7
(*Continued*)

Line	*Speaker*	*Utterance*	*Interpretive Comment*	*Mechanism*[a]	*Collaboration*
12	HD3	You see, the split here is 5 and 3, total 8 parts. But then down here, split will be 4 and 3, divide into 7 parts . . . both have to be the same total parts I think.	HD3 uses the diagram to elaborate and also to explain ratios in terms of part–whole relations	b, c	Explanation
13	HD1	I see, ah, yes . . .	HD1 finally understands		Agreement
14	HD2	Maybe we can use just one of the ratios	HD2 proposes a modification to the proposal	a	Proposal
15	HD1	No, then we will get different values . . .	HD1 demonstrates his understanding and evaluates that the modification will not work drawing attention to the non-additivity of ratios—a critical feature	b, c	Evaluation, explanation
16	HD3	I got 3.12 something and 2.86, see . . . which one to choose?	HD3 supports HD1 with calculations	b, c	Elaboration, question
17	HD2	You're right . . . we can't get it this way . . . need to think of another method	HD2 agrees that this method will not work	c	Agreement, evaluation, proposal
18	HD1	Okay, so how?	HD1 refocuses the group on the problem		Question

Note. HD = high diversity; LCM = lowest common multiple.

[a]Mechanism a = activation and differentiation of prior knowledge in relation to the targeted concepts; mechanism b = attention to critical conceptual features of the targeted concepts; mechanism c = explanation and elaboration of these features.

TABLE 8
Solution Exploration by the LD Group

Line	*Speaker*	*Utterance*	*Interpretive Comment*	*Mechanism*[a]	*Collaboration*
1	LD3	How long will she bike . . . umm . . . we don't know	LD3 identifies an unknown parameter (biking time)	a	Proposal
2	LD2	. . . We make a guess?	LD2 proposes a solution method	a	Proposal
3	LD1	Guess. . . . how?	LD1 seeks an explanation		Question
4	LD2	Yah, let's say 20 minutes? Then you see she will cover 3,000 m	LD2 explains; performs computation	c	Elaboration
5	LD3	And Hady in 20 minutes will walk. . . . 1,000	LD3 supports LD2 with further computation	c	Elaboration
6	LD2	yes	LD2 agrees		Agreement
7	LD1	but we need to find where they will change over . . .	LD1 reminds the group of the goal of the problem	c	Evaluation
8	LD2	No no no . . . slowly . . . try. . . . try 15 for Jasmine . . . 15 minutes . . . so she will cover 2,250 . . . and Hady for 15 will cover 750 meters . . .	LD2 continues with the trial and error computations	c	Disagreement, elaboration
9	LD1	I don't see how this will give us the answer	LD1 disagrees with LD2		Evaluation
10	LD2	For 30 minutes, Jasmine covers 4,500 meters and. . . . Hady 1,500	LD2 continues with the computations		Elaboration
11	LD1	Can you explain what you are doing?	LD1 seeks an explanation from LD2		Question
12	LD2	Trying to make it 5 km . . . first we got 4,000, then here 3,000 and now 6,000 . . .	LD2 explains	c	Explanation
13	LD3	Wait . . . wait . . . Jasmine's walking and biking together should be 5 km, not Jasmine and Hady . . .	LD3 draws attention to a key parameter—total distance—but does not elaborate	b	Evaluation
14	LD2	Oh yah . . . thanks	LD2 understands and agrees		Agreement
15	LD1	Maybe we guess distance because that's what the question asks us to find . . .	LD1 proposes a modification to the trial and error solution	a	Proposal
16	LD3	Maybe we can use LCM . . .	LD3 proposes another solution path	a	Proposal

(*Continued*)

TABLE 8
(*Continued*)

Line	*Speaker*	*Utterance*	*Interpretive Comment*	*Mechanism*[a]	*Collaboration*
17	LD2	LCM how?	LD2 seeks an explanation		Question
18	LD3	We find the LCM of their speeds and use it to find the distance . . .	LD3 explains	a, c	Explanation
19	LD1	I think we should just use guess and check directly . . . easier . . .	LD1 argues for trial and error with reasons	c	Disagreement, proposal
20	LD2	Why don't you do LCM and we can work on guess and check and then we see okay?	LD2 proposes that the group work on both solutions		Proposal (for division of labor)
21	LD3	Okay . . .	LD3 agrees		Agreement
22	LD1	Okay . . .	LD1 agrees		Agreement

Note. LD = low diversity; LCM = lowest common multiple.

[a] Mechanism a = activation and differentiation of prior knowledge in relation to the targeted concepts; mechanism b = attention to critical conceptual features of the targeted concepts; mechanism c = explanation and elaboration of these features.

(see the "Mechanism" columns in Tables 7 and 8). Because Group LD's discussion seemed to be focused mainly on computational features of the guess and check method, there was little evidence that members attended to the critical features of the targeted concept; that is, mechanism b was rarely invoked. In contrast, Group HD's discussion was at a more conceptual level; that is, the group worked out the ratios of the walking and biking speeds and seemed to have realized that the ratio method does not work when the walking and biking speed ratios are different. That is, unless the denominators of the ratios are the same, one cannot simply add the ratios. This non-additive property is precisely the conceptual feature that needed to be attended to, and Group HD students seemed to have been able to do that. This contrast between the two groups can be seen in the difference in the number of times mechanisms b and c seemed to have been invoked in their respective excerpts in Tables 7 and 8. Note that the guess and check method does afford opportunities for attending to the non-additive property of ratios, but Group LD seemed largely focused on computational as opposed to critical features.

In terms of collaboration, it once again seemed that there were differences between the two groups (see the "Collaboration" columns in Tables 7 and 8). In Group HD, solution proposals were met with questions, clarification and agreement, followed by evaluation and more questions, leading to explanations and evaluation and then more explanation until shared understanding was established (Utterances 4–13). In other words, collaboration seemed to have facilitated attention to and elaboration of critical features, that is, mechanisms b and c. In contrast, the collaborative pattern in Group LD was mainly one of solution proposal, followed by question, explanation and computation, with disagreements or alternative viewpoints not being taken up substantively for discussion (Utterances 2–15). Although there was evidence of explanation and elaboration, these explanations tended to describe the computational features (e.g., trying to make it 5 km) as opposed to explaining and elaborating upon the conceptual features relating to the average speed of Hady and Jasmine (Utterances 11–12). Consequently, although Group LD was able to compute and work on the trial and error method to get very close to a successful solution (see Figure 4), the excerpt reveals how a heavy emphasis on computation could have come at the expense of conceptual elaboration.

Summary. The preceding mixed-method analysis attempted to unpack variation in the generation and exploration phase of the PF design to explain how differences between PF groups related to differences in posttest performance. First, the quantitative analysis of RSM diversity in PF groups shows that the greater the number of RSMs generated by a group, the better the posttest performance of the group members (mechanism a). The fact that RSM diversity explained the most variance underscores its explanatory importance. Second, the qualitative contrasting-case analysis serves to illustrate how two groups that differed in the number of RSMs they generated additionally seemed to differ in their collaborative understanding of the problem and exploration of the RSMs and how this difference in exploration potentially influenced opportunities to attend to, explain, and elaborate upon the critical features of the concept of average speed (mechanisms b and c). However, these findings remain tentative; analysis of all PF groups needs to be carried out before any stronger claims can be made.

GENERAL DISCUSSION

This study was designed to explore the hidden efficacies, if any, in delaying structure in the learning and performance space of students by having them engage in unscaffolded problem solving of complex problem scenarios prior to DI. We carried out design experiments in three schools comprising students of significantly different general and mathematical ability.

School A comprised students of average ability as evidenced by performance on national standardized examinations, that is, the PSLE. This was the second iteration of implementation of the average speed curricular unit, and the same teacher had been involved in the project for both iterations. Consistent with findings from the previous cohort of students taught by the same teacher as the one involved in this study (Kapur, 2009), findings suggest that students from the PF condition outperformed those from the DI condition on the well-structured problem items, the complex problem item, as well as the graphical representation item on the posttest, thereby providing support for the PF hypotheses.

School B comprised students of significantly lower general and mathematical ability than those from School A, and this was the first implementation in which the school had participated. Findings from School B replicated the findings from School A on the complex and representational flexibility problem items. However, for the well-structured items, only one of the two PF classes significantly outperformed the DI class. Because the descriptive trend was similar to that in School A, that is, PF > DI, and because PF students were not given any homework assignments or intensive practice on well-structured problems, it was encouraging that they still managed to outperform DI students on the very kinds of well-structured problems on which the DI students had received intensive practice and feedback.

School C comprised students of even lower general and mathematical ability than those from School B, and like School B, this was the first implementation in this school. Because of curricular time constraints, School C presented us with an additional constraint of testing the PF hypothesis within a considerably shortened time, as a result of which we were not sure whether we could expect our hypotheses to hold in this school. Notwithstanding, although the descriptive trend of PF > DI in School C was consistent with what was found in Schools A and B, the difference reached significance only for the complex problem and representational flexibility items. Even so, findings were mixed in that only one of the two PF classes seemed to have significantly outperformed the DI class.

In sum, we want to emphasize three significant findings. First, we found that compared to DI, PF seems to engender deeper conceptual understanding without compromising performance on well-structured problems. This suggests that PF could be a pedagogically tractable design in classrooms across a spectrum of schools with different mathematical ability levels, a finding that is consistent with a growing body of classroom-based research programs (e.g., diSessa et al., 1991; Lesh & Doerr, 2003; Schwartz & Martin, 2004).

Second, although we found a significant difference among schools in terms of their students' ability to generate RSMs for solving the novel, complex problems, this difference among the schools had a notably smaller effect size ($\eta^2 = .04$) than preexisting differences in general ability ($\eta^2 = .85$) and mathematical ability ($\eta^2 = .44$) as measured by the PSLE. In other words, differences in the ability of students to generate RSMs to novel, complex problems are not as large as one would expect given the differences in general and mathematical abilities.

Third, we found that RSM diversity was correlated with learning outcomes; that is, the greater the RSM diversity, the better the learning outcomes on average. Furthermore, the effect of RSM diversity on learning outcomes far exceeded the effect of school or preexisting differences in prior knowledge (recall that the effect of RSM diversity was about 9 times stronger than the effect of pretest and 13 times stronger than that of the school).

Taken together, these findings emphasize the need to design and understand conditions under which delaying structure in learning and problem-solving activities can enhance learning (Kapur, 2008, 2009, 2010).

Explaining PF

To explain PF, we need to explain why students from the PF condition performed better, on average, than students from the DI condition. What were the kinds of processes that PF students were involved in that made for better learning and performance?

The PF design embodied four interdependent core mechanisms of (a) activation and differentiation of prior knowledge, (b) attention to critical features, (c) explanation and elaboration of these features, and (d) organization and assembly into canonical RSMs. In the PF design, the activity, the participation structures, and the social surround were designed to facilitate these core mechanisms. Evidence suggests that (a) PF groups were, on average, able to generate multiple RSMs for solving the complex problems; (b) with a few exceptions, PF groups were not successful in solving the problems; yet (c) PF students, on average, outperformed DI students on the posttest. As hypothesized, the process of generating and exploring the RSMs may have engendered sufficient knowledge differentiation and attention to critical features that in turn prepared students to better discern and understand those very concepts and RSMs when presented in a well-assembled form during the consolidation phase (diSessa et al., 1991; Schwartz & Bransford, 1998; Spiro et al., 1992). Therefore, the better performance of the PF condition provides support for the theoretical conjectures embodied in the PF design.

Further support for the core mechanisms embodied in the PF design comes from the analysis of variation within the PF condition. Quantitative analysis of RSM diversity in all PF groups across the three schools showed that the greater the number of RSMs generated, the better the learning outcome on average. This provides support for the core mechanisms (especially mechanism a) embodied in the PF design.

Building upon this, the qualitative contrasting-case analysis further illuminates how two groups that differed in the number of RSMs they generated additionally differed in their collaborative understanding of the problem and exploration of the solutions and how this difference in exploration influenced opportunities to attend to, explain, and elaborate upon the critical features of the problem and the concept of average speed. This provides support for the core mechanisms (especially mechanisms b and c) embodied in the PF design.

Finally, we acknowledge that space did not permit an analysis of the consolidation phase, and consequently there is no evidence for core mechanism d. However, this forms the thrust of our work as we continue further analysis to build the explanatory base for the PF design.

Additional explanatory conjectures. Now we consider additional explanatory conjectures for other mechanisms that emerged from our work. We acknowledge that being conjectures, they require more research to examine them further.

An affective dimension of *ownership* emerged, which is consistent with diSessa et al.'s (1991) findings. From our observations in the classrooms as well as debriefings with the teachers, it seemed that PF students exhibited strong ownership of the RSMs they developed. In future studies, we hope to unpack the role of ownership in PF further—that is, how the extent and nature of student ownership of ideas and methods influences the extent and nature of what students learn from PF experiences.

In addition to this explanation, there are conceivably other possible contributing factors that will need to be studied more closely. For example, there is some indication from the group discussions that the PF design gave students opportunities to engage and develop their meta-cognitive and self-regulatory functions, which in turn are a critical component of learning and problem-solving expertise. In contrast, in the DI design, the well-structured problems may not have afforded such opportunities. Examining the collaborative problem-solving processes to unpack the roles of meta-cognitive and self-regulatory functions in PF is an area that future studies and analysis would do well to examine further.

Yet another explanatory conjecture deals with the notion that perhaps PF students had greater opportunities to engage in the practice of mathematics (Thomas & Brown, 2007). After all, the acts of representing problems, developing domain-general and -specific methods, flexibly adapting or inventing new RSMs when others do not work, critiquing, elaborating, explaining to one another, and ultimately not giving up signify the kinds of epistemic resources that mathematicians commonly demonstrate and leverage in their practice (diSessa et al., 1991). This notion also resonates well with Brown's (2008) notion of *tinkering* as a mode of knowledge production, that is, designing learning in ways that provides opportunities to "play" with knowledge, generate ideas, share and critique, and ultimately strive to understand the effectiveness of one's ideas (Bielaczyc & Kapur, 2010). Having opportunities to engage in processes that afford such tinkering, processes that Scardamalia (2009) referred to as *epistemic invention*, may have helped students expand their repertoire of epistemic resources situated within the context of classroom-based problem-solving activities (Hammer, Elby, Scherr, & Redish, 2005).

To be clear: We are not arguing that some larger epistemic shift took place within a short design intervention. What we are arguing instead is that perhaps the PF design provided students with the opportunities to take the first steps toward developing these context-dependent, epistemic resources for tinkering (Hammer et al., 2005). The more such opportunities are designed for students, the better they may develop such epistemic resources, the greater the likelihood of better learning and performance. Once again, these remain plausible conjectures that future studies and analysis would do well to examine further.

Limitations and Future Work

It is of course too early to attempt any generalization of the claims; the scope of inference holds only under the conditions and settings of the respective study and is thus circumscribed by the content domain, communication modality, age group, sociocultural factors, and so on. Given the reality of working in real classroom contexts that rarely allow for strict causal attribution to design elements, findings from the experiments may only be attributed to the various instructional designs as wholes and not to their constituent design elements. Furthermore, although the three schools were sampled because of the significantly different general and mathematical ability of their student intakes, differences among schools are difficult to attribute to a particular factor because of multiple confounds between school, teacher, student ability, and intervention specifications.

In addition to evidencing the explanatory conjectures identified in the preceding section, we aim to continue to further expand the explanatory basis for PF by further unpacking the variation in the generation and exploration phase and the consolidation phase. For the generation and exploration phase, we aim to examine the nature of interactional behaviors and RSM sequences and relate them to gains in group and individual outcomes. Methods such as lag-sequential analysis may potentially be useful for unpacking the temporal variation in RSM sequences (e.g., Kapur, 2011). In addition, we want to examine the role of learners' motivation and frustration thresholds in learning from PF. For the consolidation phase, we aim to carry out a mixed-method analysis of teachers' orchestration of the consolidation lessons vis-à-vis the design principles to see how variation in teacher-led consolidation relates to variation in outcomes. Further work and analyses at multiple grain sizes with both students and teachers might speak to these concerns and add further explanatory power to PF.

CONCLUSION

At the heart of the work reported in this paper lies the incommensurability between learning and performance; that is, conditions that maximize performance in the shorter term *may not necessarily* be the ones that maximize learning in the longer term (Clifford, 1984; Schmidt & Bjork, 1992). Four possibilities for design emerge. First is the possibility of designing conditions that maximize performance in the shorter term and that also maximize learning in the longer term. Let us call such design efforts designing for *productive success*. Indeed, a substantial amount of research in the cognitive and learning sciences speaks to this, and rightly so, because understanding conditions under which designing structure in learning and problem-solving activities can lead to productive success is an important line of research. However, there is also the concomitant possibility of designing conditions that may well not maximize performance in the shorter term but in fact maximize learning in the longer term. Let us call such design efforts designing for *productive failure*. Consistent with past research, findings reported in this paper suggest that there are conditions under which delaying structure in learning and problem-solving activities may in fact lead to PF. Note that the proposition is not that one must always design by delaying structure to understand the conditions under which doing so may lead to PF. Instead, what is being proposed here is that as a field we stand to gain more if we engage in research that seeks to understand both PF as well as productive success and that a dual focus stands to advance the field in ways that neither single focus alone can (Kapur & Rummel, 2009). As these lines of inquiry push back against and inform each other, will we generate not only better understandings of PF and productive success but also better understandings of the other two possibilities: conditions under which designs lead to *unproductive success*—an illusion of performance without learning—as well as *unproductive failure*.

ACKNOWLEDGMENTS

The research reported in this paper was funded by a grant to Manu Kapur from the Learning Sciences Lab of the National Institute of Education of Singapore. The claims and opinions presented herein are ours alone and do not necessarily represent those of the funding agency. We would like to thank the students, the teachers, the heads of the department of mathematics, and the principals of the participating schools for their support for this project. Parts of this paper were also presented at the 2008 and 2009 annual meetings of the Cognitive Science Society. We are indebted to Leigh Dickson and June Lee for their help with the logistics and coordination with the schools. We are particularly grateful to Professor Steven Zuiker for his careful and insightful comments and suggestions on this manuscript.

REFERENCES

Ainsworth, S., Bibby, P., & Wood, D. (2002). Examining the effects of different representational systems in learning primary mathematics. *Journal of the Learning Sciences, 11*, 25–61.

Anderson, J. R. (2000). *Cognitive psychology and its implications*. New York, NY: Worth.

Barron, B. (2003). When smart groups fail. *Journal of the Learning Sciences, 12*, 307–359.

Bielaczyc, K., & Kapur, M. (2010). Playing epistemic games in science and mathematics classrooms. *Educational Technology, 50*(5), 19–25.

Brown, J. S. (2008). *Tinkering as a mode of knowledge production* [Videocast]. Retrieved from www.johnseelybrown.com

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*(1), 32–41.

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

Clifford, M. M. (1984). Thoughts on a theory of constructive failure. *Educational Psychologist, 19*(2), 108–120.

Cobb, P. (1995). Cultural tools and mathematical learning: A case study. *Journal for Research in Mathematics Education, 26*, 362–385.

Cobb, P., Wood, T., & Yackel, E. (1993). Discourse, mathematical thinking and classroom practice. In E. Forman, N. Minick, & C. Stone (Eds.), *Contexts for learning: Sociocultural dynamics in children's development* (pp. 91–119). New York, NY: Oxford University Press.

Cognition and Technology Group at Vanderbilt. (1997). *The Jasper Project: Lessons in curriculum, instruction, assessment, and professional development*. Mahwah, NJ: Erlbaum.

Cohen, E. G., Lotan, R. A., Abram, P. L., Scarloss, B. A., & Schultz, S. E. (2002). Can groups learn? *Teachers College Record, 104*, 1045–1068.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York, NY: Academic Press.

Dillenbourg, P. (2002). Over-scripting CSCL: The risks of blending collaborative learning with instructional design. In P. A. Kirschner (Ed.), *Three worlds of CSCL: Can we support CSCL* (pp. 61–91). Heerlen, The Netherlands: Open Universiteit Nederland.

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

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

Garner, W. R. (1974). *The processing of information and structure*. Potomac, MD: Erlbaum.

Gibson, J. J., & Gibson, E. J. (1955). Perceptual learning: Differentiation or enrichment? *Psychological Review, 62*, 32–41.

Greeno, J. G., Smith, D. R., & Moore, J. L. (1993). Transfer of situated learning. In D. K. Detterman & R. J. Sternberg (Eds.), *Transfer on trial: Intelligence, cognition, and instruction* (pp. 99–167). Norwood, NJ: Ablex.

Hammer, D., Elby, A., Scherr, R. E., & Redish, E. F. (2005). Resources, framing, and transfer. In J. P. Mestre (Ed.). *Transfer of learning from a modern multidisciplinary perspective* (pp. 89–120). Greenwich, CT: Information Age.

Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark (2006). *Educational Psychologist, 42*(2), 99–107.

Hox, J. J. (1995). *Applied multilevel analysis*. Amsterdam, The Netherlands: TT-Publikaties.

Jonassen, D. H. (2000). Towards a design theory of problem solving. *Educational Technology, Research and Development, 48*(4), 63–85.

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

Kapur, M. (2009). Productive failure in mathematical problem solving. *Instructional Science*, *38*, 523–550. doi:10.1007/s11251-009-9093-x

Kapur, M. (2010). A further study of productive failure in mathematical problem solving: Unpacking the design components. *Instructional Science*, *39*, 561–579. doi:10.1007/s11251-010-9144-3

Kapur, M. (2011). Temporality matters: Advancing a method for analyzing problem-solving processes in a computer-supported collaborative environment. *International Journal of Computer-Supported Collaborative Learning, 6*(1), 39–56.

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

Kapur, M., & Lee, J. (2009). Designing for productive failure in mathematical problem solving. In N. Taatgen & V. R. Hedderick (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2632–2637). Austin, TX: Cognitive Science Society.

Kapur, M., & Rummel, N. (2009). The assistance dilemma in CSCL. In C. O'Malley, D. Suthers, P. Reimann, & A. Dimitracopoulou (Eds.), *Proceedings of the Computer-Supported Collaborative Learning Conference* (pp. 37–39). Rhodes, Greece: International Society of the Learning Sciences.

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

Lesh, R. (1999). The development of representational abilities in middle school mathematics. In I. E. Sigel & K. Tyner (Ed.), *Development of mental representation: Theories and applications* (pp. 323–349). Hillsdale, NJ: Erlbaum.

Lesh, R. R., & Doerr, H. M. (2003). *Beyond constructivism: Models and modeling perspectives on mathematics problem solving, learning, and teaching*. Mahwah, NJ: Erlbaum.

Lobato, J. (2003). How design experiments can inform a rethinking of transfer and vice versa. *Educational Researcher, 32*(1), 17–20.

Nathan, M. J., & Kim, S. (2009). Regulation of teacher elicitations in the mathematics classroom. *Cognition and Instruction, 27*, 91–120.

Puntambekar, S., & Hübscher, R. (2005). Tools for scaffolding students in a complex learning environment: What have we gained and what have we missed? *Educational Psychologist, 40*(1), 1–12.

Sandoval, W. A. (2004). Developing learning theory by refining conjectures embodied in educational designs. *Educational Psychologist, 39*, 213–223.

Scardamalia, M. (2009, February). *The knowledge creation imperative*. Invited talk at the National Institute of Education, Singapore.

Scardamalia, M., & Bereiter, C. (2003). Knowledge building. In J. W. Guthrie (Ed.), *Encyclopedia of education* (pp. 1370–1373). New York, NY: Macmillan Reference.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3*, 207–217.

Schwartz, D. L. (1995). The emergence of abstract representations in dyad problem solving. *Journal of the Learning Sciences, 4*, 321–354.

Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. *Cognition and Instruction, 16*, 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*, 129–184.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 592–604.

Spiro, R. J., Feltovich, R. P., Jacobson, M. J., & Coulson, R. L. (1992). Cognitive flexibility, constructivism, and hypertext. In T. M. Duffy & D. H. Jonassen (Eds.), *Constructivism and the technology of instruction: A conversation* (pp. 57–76). Hillsdale, NJ: Erlbaum.

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences*. Mahwah, NJ: Erlbaum.

Thomas, D., & Brown, J. S. (2007). The play of imagination: Extending the literary mind. *Games and Culture, 2*(2), 149–172.

Tobias, S., & Duffy, T. M. (2010). *Constructivist instruction: Success or failure*. New York, NY: Routledge.

Van Lehn, K., Siler, S., Murray, C., Yamauchi, T., & Baggett, W. B. (2003). Why do only some events cause learning during human tutoring? *Cognition and Instruction, 21*, 209–249.

Voss, J. F. (1988). Problem solving and reasoning in ill-structured domains. In C. Antaki (Ed.), *Analyzing everyday explanation: A casebook of methods* (pp. 74–93). London, England: Sage.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 17*, 89–100.

APPENDIX A

A Complex Problem Scenario

It was a bright, sunny morning and the day of the Singapore Idol auditions. Hady and Jasmine were going to audition as a team. They were practicing at their friend Ken's house and were planning to bike to the auditions at Singapore Expo. The auditions were supposed to start at 2 p.m. and Hady and Jasmine wanted to make sure that they could make it in time.

Hady: Ken, how do we get to the Singapore Expo from here?

Ken: Well, follow this road (pointing to a map) until you reach the expressway. I usually drive at a uniform speed of 90 km/h on the expressway for about 3 minutes. After that there is a sign telling you how to get to Singapore Expo.

Jasmine: How long does it take you to reach Singapore Expo?

Ken: It normally takes me 7 minutes to drive from my house when I am traveling at an average speed of 75 km/h.

After getting the directions, Hady and Jasmine left Ken's house and biked together at Jasmine's average speed of 0.15 km/min. After biking for 25 minutes, Jasmine biked over a piece of glass and her tire went flat.

Jasmine: Oops! My tire is flat! What shall we do now? Can I just ride with you on your bike or shall we take a bus the rest of the way?

Hady: I don't think that is a good idea. My bike is old and rusty and it cannot hold both of us. Taking the bus is not a very good idea either. There is no direct bus from here to Singapore Expo, so we would have to take one bus and then transfer to another one. All the waiting for buses would definitely make us late. Do you have any money on you?

Jasmine: Let me check. . . . I forgot to withdraw money today. I only have $2.

Hady: I did not bring my wallet. I only have $1 for a drink.

Jasmine: Since we do not have enough money to take a taxi, shall we just leave our bikes here and walk?

Hady: It takes me approximately 5 minutes to walk to school, which is about 250 meters from my home. How long does it take you to walk to school?

Jasmine: It takes me about 13–15 minutes to walk to school, which is about 450 meters from my home.

Hady: No, no, no! Walking would take too much time. We will end up late. Why don't you lock up your bike and take my bike and bike ahead. Leave my bike somewhere along the route and begin walking to the audition. I will walk from here until I get to my bike and ride it the rest of the way since I can bike at a faster speed. My average biking speed is 0.2 km/min.

Jasmine: That sounds like a good idea! But how far should I ride your bike before leaving it for you and walking the rest of the way? Since we are auditioning together as a team, we have to reach there at the same time!?

How far should Jasmine ride Hady's bike so they both arrive at the audition at the same time?

APPENDIX B

Examples of Well-Structured Problems

1. The average speed of a ship for the first hour of a journey is 32 km/hr. Its average speed for the next 2 hours is 41 km/hr. Find its average speed for the whole journey.

2. Jack walks at an average speed of 4 km/hr for one hour. He then cycles 6 km at 12 km/hr. Find his average speed for the whole journey.

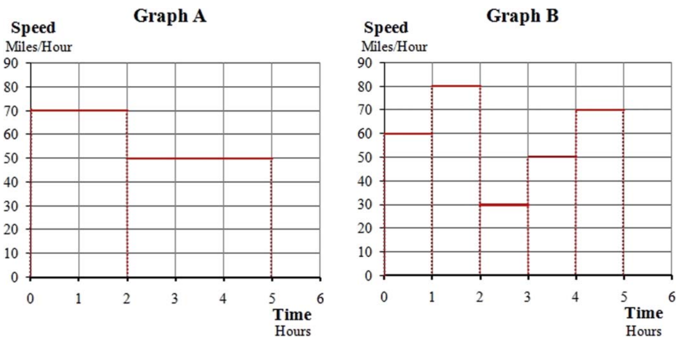
APPENDIX C

Items on the Posttest

A well-structured item. David travels at an average speed of 4 km/hr for 1 hour. He then cycles 6 km at an average speed of 12 km/hr. Calculate his average speed for the entire journey.

The complex item. Hummingbirds are small birds that are known for their ability to hover in mid-air by rapidly flapping their wings. Each year they migrate approximately 9,000 km from Canada to Chile and then back again. The Giant Hummingbird is the largest member of the hummingbird family, weighing 18–20 gm. It measures 23 cm long and it flaps its wings between 70–80 times per minute. For every 18 hours of flying it requires 6 hours of rest. The Broad Tailed Hummingbird flaps its wings 100–125 times per minute. It is approximately 10–11 cm long and weighs approximately 3–4 gm. For every 12 hours of flying it requires 12 hours of rest. If both birds can travel 1 km for every 550 wing flaps and they leave Canada at approximately the same time, which hummingbird will get to Chile first?

The graphical representation item. Bob drove 140 miles in 2 hours and then drove 150 miles in the next 3 hours. Study the two speed–time graphs A and B carefully. Which graph—A, B, or both—can represent Bob's journey?



This item was adapted from Stanford Research International's research on SimCalc and the Math of Change.