Cognition and Instruction

Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t775648096

Productive Failure

Manu Kapur a

a National Institute of Education, Nanyang Technological University, Singapore

Online Publication Date: 01 July 2008

To cite this Article: Kapur, Manu (2008) 'Productive Failure', Cognition and Instruction, 26:3, 379 — 424

To link to this article: DOI: 10.1080/07370000802212669
URL: http://dx.doi.org/10.1080/07370000802212669

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.informaworld.com/terms-and-conditions-of-access.pdf

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
Productive Failure

Manu Kapur
National Institute of Education, Nanyang Technological University,
Singapore

This study demonstrates an existence proof for productive failure: engaging students in solving complex, ill-structured problems without the provision of support structures can be a productive exercise in failure. In a computer-supported collaborative learning setting, eleventh-grade science students were randomly assigned to one of two conditions to solve problems in Newtonian kinematics. In one condition, students solved ill-structured problems in groups followed by well-structured problems individually. In the other condition, students solved well-structured problems in small groups followed by well-structured problems individually. Finally, all students solved ill-structured problems individually. Groups who solved ill-structured problems expectedly struggled with defining and analyzing the problems, resulting in poor quality of solutions. However, despite failing in their collaborative efforts, these students outperformed their counterparts in the well-structured condition on individual near- and far-transfer measures subsequently, suggesting a latent productivity in what initially seemed to be failure.

Structuring problem-solving tasks so that learners accomplish what they may not in the absence of structure is a prolific area of research in education and the learning sciences. Structure, broadly conceived, comes in a variety of forms such as structuring the problem or task itself, scaffolding, providing tools and resources, expert help, and so on. In such research, “what can be accomplished” generally refers to a problem or task that is often complex and ill-structured, and one that is beyond the existing skill set and abilities of the learner (Brown, Collins, & Duguid, 1989; Scardamalia & Bereiter, 2003). The complex and ill-structured nature of the problem thus demands, as is often argued, that support structures be provided for learners as they solve such problems, for without the support, learners may fail. However, this argument inadvertently implies that there is little...
efficacy in learners solving complex, ill-structured problems without the provision of support structures during problem solving. Although this implication may well be grounded in empirical evidence, it is also conceivable that leaving learners to struggle and even fail at tasks that are ill-structured and beyond their skills and abilities may in fact be a productive exercise in failure. It is this possibility that is explored in this article; the challenge for research lies in conceptualizing ways of extracting the hidden efficacies in the seemingly unproductive production that often results, at least in the shorter term, when learners engage in solving complex, ill-structured problems. In this article, I demonstrate an existence proof for productive failure in an instructional setting using empirical evidence from a large-scale study of computer-supported collaborative learning (CSCL) groups.

FAILURE AND STRUCTURE

The role of failure in learning and problem solving is no doubt intuitively compelling. Research on impasse-driven learning in coached problem-solving situations provides strong evidence for the role of failure in learning, in general, and in problem solving, in particular (VanLehn, 1999; VanLehn, Siler, Murray, Yamauchi, & Baggett, 2003). Examining coached problem-solving episodes for events that result in learning, Van Lehn et al. (2003) found that not only was successful learning of a principle (e.g., a concept, a natural law in physics) associated with events when students reached an impasse during problem solving but also that when students did not reach an impasse, learning was rare despite explicit tutor-explanations of the target principle. Instead of providing immediate structure in the form of feedback, questions, or explanations when the learner demonstrably makes an error or is “stuck,” their 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 this delaying of structure in the context of text comprehension, McNamara, Kintsch, Songer, and Kintsch (1996) found that whereas low-knowledge learners tended to benefit from high-coherence texts, high-knowledge learners benefited from low-coherence texts, and especially more so when a low-coherence text preceded a high-coherence one (McNamara, 2001). This, McNamara argues, suggests that reading low-coherence texts may force learners to engage in compensatory processing by using their prior knowledge to fill in the conceptual gaps in the target text, in turn, preparing them better to leverage a high-coherence text subsequently. Further evidence for such preparation for future learning (Schwartz & Bransford, 1998) can be found in the inventing to prepare for learning (IPL) research by Schwartz and Martin (2004). In a sequence of design experiments on the teaching of descriptive statistics, Schwartz and Martin (2004) demonstrated a hidden efficiency of invention activities when they preceded direct instruction
(e.g., lectures), although these activities failed to produce canonical conceptions and solutions during the invention phase.

Clearly, the relationship between failure and structure forms a common thread through these diverse research programs. It is reasonable to reinterpret the central findings collectively as an argument for a delay of structure in learning and problem-solving situations, be it in the form of feedback and explanations, coherence in texts, or direct instruction. Indeed, these studies point to the efficacy of learner-generated processing, conceptions, and understandings, even though such conceptions and understandings may, in the shorter term, not be correct and the process of arriving at them not as efficient. Even in non-instructional, design settings (e.g., engineering, architectural design), which tend to be inherently complex and ill-structured, the importance of understanding the nature of the design problem space through divergent, open-ended exploration is found to be an invariant across design situations (Goel & Pirolli, 1992). However, the aforementioned studies in instructional settings deal with students engaged in solving well-defined, well-structured problems. Furthermore, they typically deal with individual problem-solving situations, coached or otherwise. Although the IPL study (Schwartz & Martin, 2004) had students work collaboratively, it was not designed to analyze the interactional dynamics, calling instead for future research to link characteristics of interactional dynamics with subsequent gains in problem solving.

Therefore, there is a clear need to extend this important line of inquiry to instructional settings where students are engaged in solving complex, ill-structured problems. Because situative, socio-constructivist theories of learning emphasize the need to engage learners in complex, ill-structured problems and tasks, such an extension becomes particularly important (Brown et al., 1989). The role of collaboration is also central to the socio-constructivist perspective and, therefore, extension to collaborative problem-solving settings also becomes important (Stahl, 2005), particularly those that are supported by technology since technology is increasingly an important mediator of social participation and collaboration (Scardamalia & Bereiter, 2003). Finally, extensions across content topics, age groups, and cultural contexts only add greater richness, diversity, and value to the wider ecology of education and learning sciences research. Over the past two decades or so, a body of research that has sought to do this is computer supported collaborative learning (CSCL), and it is within this body of research that the present study is specifically situated. Consequently, a quick examination of CSCL research is what I focus on next.

STRUCTURE AND CSCL

In CSCL research, generally speaking, research focuses more on what is gained from structuring but not as much on what is lost. Structure, again, has been
conceived broadly in CSCL research as well. It takes on a variety of forms. I give a few examples (also see Koschmann, Suthers, & Chan, 2005): task or problem structuring (e.g., Jonassen & Kwon, 2001; Kapur & Kinzer, 2007); meta-cognitive support through reflection prompts (e.g., Lin, Hmelo, Kinzer, & Secules, 1999); content support (e.g., Fischer & Mandl, 2005); interactional support through question prompts (e.g., Ge & Land, 2003); supporting group discourse through argumentation tools (e.g., Cho & Jonassen, 2002) and representational guidance (Suthers & Hundhausen, 2003); scripting inter-dependencies through a division of labor (e.g., Schellens, Van Keer, Valcke, & De Wever, 2005); supporting the problem-solving process through process scaffolds (e.g., Schwartz, Lin, Brophy, & Bransford, 1999), and so on. Regardless of the type of structure, it is provided to support learners while they are engaged in solving a particular problem. The essential argument is that structure helps learners accomplish what they might not otherwise be able to in the absence of structure. Much empirical work and analysis supports this and has consistently shown that minimally structured problem-based learning rarely leads to anything meaningful (Kirschner, Sweller, & Clark, 2006). Hence, there is a deeply ingrained maxim that external support structures need to be provided for learners when they engage in solving complex, ill-structured problems, for without such support, they may fail.

However, this maxim does not necessarily imply that there is little or no efficacy embedded in the resulting “failure.” Why? There are at least three reasons. First, it is one thing to (arguably) infer learning from observed success on measures of performance. But the conclusion that a lack of success on those measures implies a lack of learning does not logically follow. Second, one is limited by the validity and scope of the measures of performance one adopts (Chatterji, 2003). Third, it is also reasonable to argue that external support structures may create a lock-in that restricts a fuller exploration of the problem and solution spaces (Reiser, 2004). Although this lock-in may be effective in helping learners accomplish the task efficiently, such learning may not be sufficiently flexible and adaptive in the longer term, especially when learners are faced with novel problems (Reiser, 2004; Schwartz & Martin, 2004). On the other hand, without such a lock-in, learners may explore, struggle, and even fail at solving ill-structured problems. The process may well be less efficient in the shorter term but it may also allow for learning that is potentially more flexible and adaptive in the longer term. Persisting with such a process may engender increasingly high levels of complexity in the exploration of the problem and solution spaces. In turn, this build-up of complexity may allow for learning that is potentially more flexible and adaptive (Kauffman, 1995). Taken together, the three aforementioned reasons give us reason to believe that there may well a hidden efficacy in students solving ill-structured problems without any support or scaffolds even if it seems to lead to failure in the shorter term. The challenge lies in extracting this efficacy.
Schwartz and Bransford’s (1998) contrasting-cases design may be leveraged to extract the hidden efficacy of having learners solve ill-structured problems without the provision of external support structures. Schwartz and Bransford (1998) showed that having students examine the similarities and differences among contrasting cases representing a target concept prepared them to derive greater benefit from a subsequent lecture or reading on that concept. In fact, the contrasting-case design is a special case of a more general argument from discernability on the nature of learning and transfer (Garner, 1974; Marton, 2006; Mestre, 2005; Schwartz & Bransford, 1998).

Marton (2006) argues that while traditional conceptions of transfer focus on the sameness between learning and performance situations, it is also important to focus on the differences so that by systematically varying the differences one can design for discernability to enhance learning in ways that may not be possible through a focus on sameness alone. To support his argument, Marton (2006) brought to bear evidence from several studies in different domains, including studies on the learning of Cantonese, which illustrate the gist of the argument from discernability particularly well (Ki & Marton, 2003, 2005). Cantonese words are distinguishable both by sound as well as tone. To a novice learning the language, the difference between sound and tone is initially not discernable unless another word with the same sound but a different tone follows the first. Only upon hearing the second word does the novice hear the tone in the second word (due to the difference/contrast), and discern how the tone differs from the tone in the first word. Hearing the first word influences the second because of the contrast, which retrospectively influences “hearing” of the first, because of the same contrast. According to Marton, “although the sameness of the sounds across the two words was a necessary condition for discerning the tone, it was the difference—and not the sameness—that was attended to, discerned, and transferred” (p. 529).

This study used Marton’s (2006) argument from discernability to demonstrate an existence proof for productive failure. In brief (the research design, instrumentation, and procedures are described in detail in the following section), students were randomly assigned to one of two conditions. In one condition, students solved ill-structured problems in groups followed by well-structured problems individually. In the other condition, students solved well-structured problems in groups followed by well-structured problems individually. All ill- and well-structured problems in the two conditions targeted the same content in Newtonian kinematics. Thus, students in the first condition received a problem-solving sequence in which the target content was held constant but the level of problem structuredness was varied (ill-structured problems followed by well-structured problems). Students in the second condition received a problem-solving sequence where both the level of problem-structuredness and target content were held constant. Finally, all students solved ill-structured problems individually that required them to use more advanced concepts in addition to those required to solve the preceding problems.
Based on the argument from discernability, it was hypothesized that the contrast in the level of problem-structuredness (ill-structured problems followed by well-structured problems) received by students in the first condition would help them discern how to structure and solve an ill-structured problem better. At the same time, it may also help them solve the well-structured problem that provides the contrast better. In turn, this may help them become better solvers of both well- and ill-structured problems. Thus conceptualized, the contrast between ill-structured problems followed by well-structured problems may facilitate a spontaneous transfer of problem-solving skills, which, in the absence of the contrast, may remain unrealized. Note that this contrasting-case design operates at a higher level across separate problem-solving activities, as opposed to operating within each of them (e.g., research cited earlier in this article). Furthermore, the design (ill-followed by well-structured) is also consistent with the delay of structure finding in previous research (McNamara, 2001; Schwartz & Martin, 2004; VanLehn et al., 2003), thereby setting up conditions for testing the productive failure hypothesis.

PURPOSE

The purpose of this study was to test the hypothesis of productive failure: whether or not there is a hidden efficacy in the unscaffolded, problem-solving efforts of groups of learners solving ill-structured problems and if this efficacy can be extracted using a contrasting-case design. This was done through a study of the problem-solving efforts of small groups of students in a synchronous CSCL environment.

METHOD

Participants

Participants were \( n = 309 \), eleventh-grade science students (197 male, 112 female) from 7 co-educational, English-speaking high schools in the National Capital Region of India. Students in the science stream typically study mathematics, physics, chemistry, and English as their main academic subjects. The proportion of males to females in this sample is considered typical for the science stream in the senior secondary years (eleventh and twelfth grades) in India. All seven schools shared the same curriculum prescribed by the Central Board of Secondary Education (CBSE) of India. The participants typically came from middle- to upper-middle-class families and were technologically savvy. More importantly, using data from the tenth-grade CBSE national standardized test scores in science,\(^1\)

\(^1\)Participants in this study were half-way through their eleventh grade; their tenth-grade CBSE national exams provided the most recent standardized test scores for comparison.
a one-way ANOVA did not find any significant difference between the seven schools in terms of student ability in science, $F = 1.677, p = .126$. The study was designed to reflect the schools’ mathematics and science curricula. Prior to the study, all students had completed the curricular unit on Newtonian kinematics, the targeted conceptual domain of the study.

At this juncture, a couple of contextual factors need to be highlighted to better situate this study. First, problem solving is an integral component of the curricula especially in light of the high-stakes competitive entrance examinations for the top universities in India. However, much problem-solving centers on well-structured problems ranging from simple to very difficult. Second, a cultural disposition of these students towards argumentation, often for its own sake, is also noteworthy (Sen, 2005). Together, the heavy emphasis on problem solving in a high-stakes competitive testing environment coupled with the cultural disposition toward argumentation form critical sociocultural, contextual factors within which this study and its findings ought to be interpreted.

Research Design and Procedures

A randomized experimental design was used and replicated in each of the seven participating schools. Within each school, participants were first randomly grouped into triads, resulting in $n = 103$ groups. These groups were then randomly assigned to one of two conditions to solve either ill-structured problems (50 groups) or well-structured problems (53 groups).

Groups solving ill-structured problems (hereinafter referred to as IS groups) were asked to solve two ill-structured problems without the provision of any external support structures or scaffolds. They were given the ill-structured problem scenarios and then were left to their own devices to discuss and solve the problems. Groups solving well-structured problems (hereinafter referred to as WS groups) were given the same problems but in a more structured format (described later) (Jonassen, 2000; Voss, 1988). All problems dealt with car-accident scenarios requiring students to apply concepts in Newtonian kinematics, were content validated by physics teachers, and pilot tested (for problem design and validation, see the following section). Each group solved two ill- or well-structured problems (their order counter-balanced) as appropriate to their assigned condition. No help or support was provided to any group during problem solving. This research design allowed for a comparison of a traditional problem-solving of well-structured problems with problem-solving of ill-structured problems, and more importantly, one that was not supported or structured via external scaffolds.

The study was carried out in three phases (see Table 1). In phase 1, 3 days before group work, all participants individually took a 25-item multiple-choice pre-test on concepts in Newtonian kinematics ($Cronbach’s \alpha = .74$). Appendix A presents...
In light of the study’s productive failure hypothesis, it is important to point out that the experimental manipulation lies squarely in the level of structuredness of the problem scenario—the task given to the groups. The study was designed to test if there is a hidden efficacy in the unscaffolded, problem-solving efforts of groups of learners solving ill-structured problems even though such efforts may seemingly lead to failure in the shorter term. To this end, the decision to choose problem structuring as the form of experimental manipulation of structure needs further explanation. After all, structure could have been imposed in a variety of ways as indeed the cited examples from previous CSCL research suggest. However, the reality of doing large-scale classroom-based research across seven schools is that one has to strictly adhere to the curriculum and the time allotted for the targeted curriculum unit; the productive failure hypothesis had to be tested in a relatively short period of curriculum time. Thus, it was not feasible to expect the schools to

### TABLE 1
The Three-Phase Research Design

<table>
<thead>
<tr>
<th>(Phase 1) (Individual)</th>
<th>(Phase 2) (Group)</th>
<th>(Phase 3) (Individual)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-test</td>
<td>Solve 2 well-structured problems in a counter-balanced order</td>
<td>Well-structured problems post-test</td>
</tr>
<tr>
<td>Well-structured condition</td>
<td>Ill-structured problems post-test</td>
<td>Ill-structured condition</td>
</tr>
<tr>
<td>Solve 2 ill-structured problems in a counter-balanced order</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

three sample items. Phase 2 of the study involved group work, which was carried out in the schools’ computer laboratories, where group members communicated with each other only through synchronous, text-only chat. The chat application automatically archived the transcript of their discussion and group solutions, which significantly eases data collection and is another reason for choosing CSCL settings for this study. Groups were given 1.5 hours per problem. Based on the pilot study with the previous cohort of students, the time allocation was found to be sufficient. In phase 3, all participants individually took a paper-and-pencil test comprising well-structured problems (post-test 1), creating a contrast for participants from the IS groups. Finally, all participants individually took a paper-and-pencil test comprising ill-structured problems (post-test 2). Both post-tests dealt with two car accident scenarios each, and were content validated as well. The well-structured problems in post-test 1, for which a maximum of 1.5 hours were given, targeted the same content in Newtonian kinematics as the group problems. The ill-structured problems in post-test 2 required participants to apply more advanced concepts in Newtonian mechanics. A maximum of 2 hours were given for post-test 2.
undertake major changes in curriculum and teacher training for this project. One had to seek small changes or manipulations that could likely test the hypothesis. Given the heavy emphasis on problem-solving in the curriculum, manipulating the level of structure in the problems constituted one such small change and was consequently a natural choice, especially because problem type has been shown to be a critical factor influencing interactional dynamics and group performance (Jonassen, 2000; Kapur & Kinzer, 2007; Lam, 1997). One could have sought bigger changes by manipulating structure in other ways, for example, teachers’ in-situ support and scaffolding of groups, providing more sophisticated CSCL tools with built-in support structures (e.g., argumentation tools, prompts, representational guidance, process scaffolds, and so on) than what the schools had, but this would have meant significant undertakings in terms of teacher training and technology development, testing and deployment. From a research design standpoint, it would have also made controlling for teacher effects across the classrooms and schools highly problematic, if not impossible. Instead, by choosing to work within the schools’ existing social and technological infrastructure (Bielaczyc, 2006), it is reasonable to argue that, if the productive failure hypothesis could be demonstrated with minimal changes to the school curriculum, teacher training, and technological resources, within a relatively short timeframe, then the findings speak well of the study’s practical significance.

Design and Validation of Problem Scenarios

The design of ill- and well-structured problem scenarios was closely aligned to Voss’ (1988, 2005) and Jonassen’s (2000) design typologies for problems. Accordingly, ill-structured problems were designed such that they:

- possessed many problem parameters with some unknown, or known only with some degree of confidence,
- possessed problem parameters that interacted with each other in interesting ways such that the effect of each could not be examined in isolation,
- possessed multiple solutions and solution paths,
- possessed multiple criteria for evaluating solutions, and
- required learners to make assumptions, judgments, and express personal opinions or beliefs.

After designing the ill-structured problem scenarios, their well-structured counterparts were designed by reducing the degrees of freedom in the ill-structured problem scenarios. As a result, the well-structured problem scenarios:

- possessed fewer problem parameters than their ill-structured counterparts,
- presented problem parameters to the learners with greater degree of confidence,
required the application of a limited number of regular rules and principles that were organized in predictive ways, and
had knowable, comprehensible solutions where the relationship between decision choices and all problem states was probabilistically close to being known.

Collaborative Phase Problem Scenarios

Four problem scenarios, two well- and two ill-structured, were developed for the collaborative phase (phase 2) of this study (see Appendix B1 for an ill-structured problem and its corresponding well-structured problem). The problem scenarios were aligned to the curricular objectives of the CBSE physics curriculum, which the participant schools followed. However, the problem scenarios were intentionally designed to be beyond the skills and abilities of the participants. This was ascertained through a pilot study conducted with the previous cohort of eleventh-grade science students (n = 60) from one of the participant schools. At a first reading of the scenarios, it may seem that the only difference between the two scenarios is that the ill-structured problem is a more elaborate version of the well-structured problem or that it contains additional irrelevant information. Although this is partly true, the difference is in fact more substantive. Because this difference between the well-structured and ill-structured problems is central to the study’s experimental manipulation, it needs further elaboration.

Clearly, the ill-structured problem contains many more parameters than the well-structured problem. For example, parameters such as weight, age, traffic conditions, prior violations, alcohol screening test, and so on are not in the problem space of the well-structured problem. In this sense, the ill-structured problem is indeed a more elaborate story problem. However, the additional parameters are not all completely irrelevant. They vary; some being more relevant to the problem than others as one would expect in a complex, authentic scenario (Goel & Pirolli, 1992; Voss, 1988). For example, the driver’s weight, age, and prior violations are perhaps not as relevant as traffic conditions or the alcohol screening test. Complicating the matter further, not all the parameters in the ill-structured problem are known with or specified to a high degree of certainty, requiring participants to either deduce or rely on assumptions, opinions, or beliefs. For instance, the coefficient of friction is given as a range in the ill-structured problem together with a statement about the bad road conditions in the city. This is in contrast to the well-structured problem where the coefficient of friction is set at 0.6 without any further qualification. Because of this, the ill-structured problem also allowed for greater learner agency to propose and modify parameters in the problem statement. For example, some IS groups argued and doubted the mechanic’s estimate for the coefficient of friction given the general road conditions. Finally, there are a greater number of interactions
between the parameters in the ill-structured problem than in the well-structured problem. Interactions do not allow the parameters to be considered in an isolated, additive manner, making the ill-structured problem significantly more complex than the well-structured problem (Voss, 2005). For example, it is easy to take the stopping distance to be 15 meters (i.e., the length of skid marks as stated in the well-structured problem). However, the information (provided by the mechanic’s account) of the wear and tear and the status of the braking fluid interacts with the length of the skid marks in the sense that the skid marks may not directly correspond to the stopping distance. With more wear and tear and the braking fluid running out, the stopping distance is likely to be greater than the length of the skid marks. How much longer is again unknown, which is precisely an example of the complexity and lack of structure that ill-structured problems were designed to engage students in. Thus, the varying levels of parametric relevancy and specification together with greater number of interactions between parameters made the ill-structured problems much more complex in comparison to the well-structured problems. Consequently, relative to the well-structured problem, the ill-structured problem admitted many more solutions and solution paths as well as criteria for evaluating those solutions and solution paths (Jonassen, 2000; Voss, 2005). Note, however, that the problems were similar in their respective goals: both problems required learners to take on the same role, that of a lawyer, and come to an evidence-based decision.

Four additional problem scenarios, two well- and two ill-structured, were developed for the individual post-tests 1 and 2, respectively, that is, for the phase 3 of the study (see Appendix B2 for one well-structured and one ill-structured problem).

Post-Test 1: Well-Structured Problem Scenarios

Post-test 1 comprised of two well-structured problem scenarios. The well-structured problems on post-test 1 targeted the same content in Newtonian kinematics as the collaborative phase problem scenarios (both well- and ill-structured). In other words, the well-structured problems on post-test 1 did not contain additional content that somehow privileged students from either the WS or IS groups because the content targeted in the collaboration phase problem scenarios and those on post-test 1 was the same. Conversely, neither the well- nor ill-structured problem scenarios in the collaboration phase contained additional content that could have privileged one group of students over the other in solving the post-test 1 problem scenarios. In either case, students had to use the same set of parameters and laws of kinematics to present the best case possible in support of their decision. The well-structured problems on post-test 1, however, functioned as a contrast to the ill-structured problem scenarios that the IS groups
solved; the contrast being in the level of structuredness of the problem while keeping the target content the same. By keeping the target content the same and varying the level of problem structuredness was hypothesized to help IS group students in discerning how to structure an ill-structured problem. Thus, the well-structured problems on post-test 1 can be seen as a measure of near transfer, in terms of the concepts and skills required to solve them.

Post-Test 2: Ill-Structured Problem Scenarios

Post-test 2 was comprised of two ill-structured problem scenarios that required students to use more advanced concepts in addition to those required to solve the problems in the collaborative phase. These included laws governing impact and conservation of momentum. Neither the WS nor the IS group participants encountered these extension concepts in the collaboration phase. Thus, it is reasonable to posit that exposure to the target content during the collaboration phase could not have privileged one group over the other. However, exposure to the contrast received by the students from the IS groups, who solved ill-structured problems collaboratively followed by well-structured problems (on post-test 1) individually, was hypothesized to help them discern how to structure an ill-structured problem thereby facilitating a spontaneous transfer of problem-solving skills, which, in the absence of the contrast, might have remained unrealized. Hence, performance on post-test 2 was conceivably a measure of far transfer albeit still within the domain of Newtonian kinematics.

Validation of Collaborative Phase and Post-Test Problem Scenarios

Instrument validation of all the problem scenarios was achieved in multiple ways. First, two physics teachers from the schools with experience in teaching the subject at the senior secondary levels (eleventh and twelfth grades) helped content validate the problems. Second, a senior-secondary English language teacher from one of the schools as well as the top three tenth-grade students from the same school assessed the problem statements for language and readability. This was done to ensure that language and readability were not confounding factors. Third, problem classification validation was also undertaken by having the same three students and the two physics teachers classify the problems into categories. Their classification was consistent with the researcher’s. Finally, all the problem scenarios were tested in a pilot study with a previous cohort of eleventh-grade science students (n = 60) from one of the participant schools. At each stage, feedback from the teachers and students was incorporated to make the necessary changes to the problem scenarios. The pilot study also informed the time allocation for group and individual tasks. This was done to ensure that insufficient time was not a confounding factor for the
differences between groups solving well- and ill-structured problems; they were able to complete the tasks and the time stamp in the chat environment indicated that groups tended to make full use of the allotted time. Also, requests for extra time from groups were few and far between to be of any significance. Thus, the time taken for all group or individual tasks was held constant.

Data Coding

Quantitative Content Analysis (QCA) (Chi, 1997) was used to segment and code utterances. The unit of analysis was semantically defined as the function(s) that an intentional utterance served in the problem-solving process. Bransford and Nitsch (1978) support the case for semantically defined units by viewing meaning-making and understanding as functions of the interdependence between interaction and context. They argued that to fully comprehend a given interaction, one must not only understand its words and the sentences (syntactic features), but also how it is situated in a discussion context; the meaning of an interaction not only clarifies the context but also gets clarified by the context. Thus, every utterance was segmented into one or more interaction unit(s), and coded into categories adapted from the Functional Category System (FCS)—an interaction coding scheme developed by Poole and Holmes (1995). Accordingly, each interaction unit was coded into one of seven categories:

1. Problem Analysis (PA): Statements that define or state the causes behind a problem (e.g., "I think the man was driving too fast"),
2. Problem Critique (PC): Statements that evaluate problem analysis statements (e.g., "how can you be sure that the man was driving fast"),
3. Orientation (OO): Statements that attempt to orient or guide the group’s process, including simple repetitions of others’ statements or clarifications; Statements that reflect on or evaluate the group’s process or progress (e.g., "let’s take turns giving our opinions"),
4. Criteria Development (CD): Statements that concern criteria for decision making or general parameters for solutions (e.g., "we need to find the initial speed of the car"),
5. Solution Development (SD): Suggestions of alternatives, ideas, proposals for solving the problem; Statements that provide details or elaborate on a previously stated alternative. They are neutral in character and provide ideas or further information about alternatives (e.g., "use the 2nd equation of motion"),
6. Solution Evaluation (SE): Statements that evaluate alternatives and give reasons, explicit or implicit, for the evaluations; this also included statements that simply agreed or disagreed with criteria development or solution
suggestion statements; Statements that state the decision in its final form or ask for final group confirmation of the decision (e.g., “yes, but how do we get acceleration”), or

7. **Non-Task (NT):** Statements that do not have anything to do with the decision task. They include off-topic jokes and tangents (e.g., “let’s take a break!”).

After an initial training phase, two trained doctoral students independently coded the interactions with an inter-rater reliability (*Krippendorff’s alpha*) of .84. The researcher and a physics teacher independently rated the quality of all group solutions as well as the individual post-test performances of all participants. Raters were blind to the treatment conditions. *Krippendorff’s alphas* of .86, .92, and .88 were achieved for rating group solutions, well-structured problems post-test 1, and ill-structured problems post-test 2, respectively.

**Summary of Data Sources and Measures**

Before describing the data analysis procedures and methods, it is worthwhile to summarize the data sources as well as the operationalization of the various measures presented in the previous sections (see Table 2).

**Data Analysis**

Data analysis was carried out at the group level first followed by hierarchical linear modeling and content analysis of individual performance on the post-tests. The purpose of the group-level analysis was to understand differences between the WS and IS groups in terms of:

1. the functional content of their discussions,
2. the sequential communicative patterns in their discussions,
3. the level of convergence in their discussions, and
4. the quality of solutions they produced as a group; that is, group performance.

The analysis of functional content provides an initial sense of “what” the groups discussed. Although such “coding and counting” analysis is common in CSCL research (Suthers, 2006), it does not provide any indication or measure of interactional complexity, for example, the kinds of significant interactional sequences and patterns that emerged in the group discussions. This study proposes two such indicators: the analysis of sequential communicative patterns using lag-sequential analysis as one measure; the analysis of convergence in group discussions as another. Together with the analysis of groups’ solutions, the four measures formed group-level descriptions for the differences between IS and WS groups.
TABLE 2
Summary of Data Sources and Measures

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Derived Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>i. Participant particulars indicating name, gender, and school</td>
<td>• School Affiliation</td>
</tr>
<tr>
<td></td>
<td>• Gender</td>
</tr>
<tr>
<td>ii. Tenth grade national CBSE standardized test scores in Science and</td>
<td>• Individual Science Ability (individual Science score)</td>
</tr>
<tr>
<td>English for each participant</td>
<td>• Group English proficiency (Mean English score of the group)</td>
</tr>
<tr>
<td>iii. Pre-test performance score for each participant</td>
<td>• Individual prior knowledge (individual pre-test score)</td>
</tr>
<tr>
<td></td>
<td>• Group prior knowledge (Mean pre-test score of the group)</td>
</tr>
<tr>
<td>iv. Automatically archived transcripts containing the problem-solving</td>
<td>• Functional content of interactional activity (proportion of interactional</td>
</tr>
<tr>
<td>interactions of groups, including the solutions to the problems (two</td>
<td>activity in the FCS categories)</td>
</tr>
<tr>
<td>transcripts per group because each group solved two problems)</td>
<td>• Problem-solving sequence (Lag-sequential analysis of significant transitions)</td>
</tr>
<tr>
<td></td>
<td>• Convergence (emergent modeling using a 1-dimension Markov walk)</td>
</tr>
<tr>
<td></td>
<td>• Group Performance (rated quality of group solutions)</td>
</tr>
<tr>
<td>v. Post-test 1 (well-structured problems) performance score for each</td>
<td>• Well-structured problem-solving ability (performance score on post-test 1)</td>
</tr>
<tr>
<td>participant</td>
<td></td>
</tr>
<tr>
<td>vi. Post-test 2 (ill-structured problems) performance score for each</td>
<td>• Ill-structured problem-solving ability (performance score on post-test 2)</td>
</tr>
<tr>
<td>participant</td>
<td></td>
</tr>
</tbody>
</table>

The purpose of the hierarchical linear modeling was to understand differences in the group-to-individual transfer effects between participants from the WS and IS conditions. As reported earlier, the transfer effects were measured through individual performance on well-structured problems (near transfer) first, followed by that on ill-structured problems (far transfer). Content analysis of students’ solutions was also carried out to provide some explanatory support for the findings.

Given the extensive nature of the group- and individual-level analyses, the variables and procedures used in the data analyses are described together with the results in the following section. It is important to note that in all the results reported in this article—at the group and the individual levels—the effects of confounding factors (e.g., school, gender, counter-balanced problem order) and
covariates (e.g., individual pre-test score, group prior knowledge as measured by mean pre-test score) were controlled for. Recall that time-on-task, both at the group and individual levels, was held constant by design.

RESULTS AND DISCUSSION

Group-Level Analysis 1: Functional Content

Controlling for the effects of school, counter-balanced problem order, group prior knowledge, and group English proficiency, a MANCOVA (with proportion of interactional activity in the six functional categories PA, PC, OO, CD, SD, and SE as the six dependent variables) revealed a significant multivariate effect of problem type (well- vs. ill-structured) on the functional content, $F = 4.91, p < .001$, partial $\eta^2 = .24$, power = .99. The Box’s test for violations of homogeneity was not significant, $F = .95, p = .521$, as were all six univariate Levene’s test for equality of error variances. Table 3 presents the descriptive statistics.

Univariate analyses showed that IS groups had significantly greater proportion of activity centered on:

- PA: problem analysis, $F = 18.20, p < .001$, partial $\eta^2 = .16$,
- PC: problem critique, $F = 11.91, p = .001$, partial $\eta^2 = .11$, and
- CD: criteria development, $F = 4.09, p = .046$, partial $\eta^2 = .04$.

Even though grade 1 onward, the medium of instruction in the participating schools is English, English is not the native language. Because students had to communicate with each other using text-only chat, English proficiency was included in the analysis to capture any variation that may be explained by differences in English proficiency.

The sum of proportions, by definition, equals one. Hence, it introduces linear redundancy in multivariate analysis. Therefore, the proportion of NT activity was excluded.

As a rule of thumb, partial $\eta^2 = .01$ is considered a small, .06 medium, and .14 a large effect size (Cohen, 1977).
In contrast, WS groups had significantly greater proportion of activity centered on:

- SD: solution development, $F = 7.23$, $p = .008$, partial $\eta^2 = .07$, and
- SE: solution evaluation activity, $F = 7.92$, $p = .006$, partial $\eta^2 = .07$.

There was no significant difference in the orientation (OO) activity between WS and IS groups, $F = .025$, $p = .874$.

**Group-Level Analysis 2: Sequential Communicative Patterns in Group Discussion**

Lag-sequential analysis\(^5\) (LSA) treats each interactional unit (defined earlier) as an observation; a coded sequence of these observations forming the interactional sequence of a group discussion (Erkens, Kanselaar, Prangsma, & Jaspers, 2003). It detects the various non-random aspects of interactional sequences to reveal how certain types of interactions follow others more often than what one would expect by chance (Wampold, 1992). It accomplishes this by identifying statistically significant transitions from one type of interactional activity to another (Bakeman & Gottman, 1997; Wampold, 1992).

For the present study, LSA revealed significant differences between the discussions of WS versus IS groups (see Figure 1). A circled category means that groups in that condition were at least twice as likely to sustain that type of activity (i.e., the activity was at least twice as likely to appear in coherent clusters rather than be spread throughout the discussion). For example, WS groups were at least twice as likely to sustain problem analysis (PA) activity. In other words, PA was at least twice as likely to be followed by more PA in WS groups than in IS groups. An arrow represents a directed transition. Contrastingly in IS groups, for example, PA activity was at least twice as likely to be followed by PC activity.

Figure 1 suggests that, with regard to how groups sustained different types of activities, IS groups were at least twice as likely to sustain PC, SE, and NT activities. For example, sequences where PC was followed by PC, and inductively, more PC, were twice as likely to be found in IS group discussions than in WS group discussion. In contrast, WS groups were at least twice as likely to sustain PA, CD, and SD activities. Note how structuring the problem reproduced the patterns of interaction that process scaffolds typically engender, that is, helping groups carry out PA, CD, SD activities in coherent phases. This lends further credence to problem structuring as a kind of an external imposition of structure on the learning and performance space. With regard to transitions, SD-SE transition

---

\(^5\)The software program Multiple Episode Protocol Analysis (MEPA) developed by Gijsbert Erkens was used for carrying out the LSA. See [http://edugate.fss.uu.nl/mepa/index.htm](http://edugate.fss.uu.nl/mepa/index.htm).
was the only significant transition that WS groups were more likely to exhibit. In contrast, the discussions of IS groups were more likely to exhibit many significant transitions (PA-PC, PA-CD, and CD-SD) as well as feedback loops (SE-PA and SE-PC, completing the loop between SE, PA, and PC).

Taken together, the discussions of WS groups were more likely marked by interactional sequences of PA-PA-PA, CD-CD-CD, SD-SD-SD (3 instances are chosen to denote the notion of sustained activity), and SD-SE. Discussions of IS groups, by contrast, were more likely marked by sequences such as PA-PC-PC-PC, PC-PC-PC, PA-CD, PA-CD-SD, CD-SD, SE-SE-SE, SE-PA, SE-PA-PC, SE-PA-PC-PC, SE-PA-CD, SE-PA-CD-SD, SE-PC, and SE-PC-PC-PC. This suggests a greater complexity in the interactional dynamics of IS than WS groups. This is because the greater the number of significant transitions and feedback loops, the greater the number of possibilities in which the discussion could unfold from any given point in the discussion. Of course, an intuitive way of understanding this is to realize that the greater the number of interactions between the components (functional categories) of a given system (group discussion), the greater is its complexity. This measure was derived from Kauffman’s (1995) measure of complexity for the evolution of Boolean networks.6

6The derivation becomes clearer when one conceives the functional categories as component states of an evolving Boolean network; Boolean in the sense that, at any point in time, a component state (PA, PC, etc.) may be present or absent in the group discussion, and interactions between the component states may be represented in terms of probabilistic logical functions. This, in many ways, is what LSA attempts to do; it looks at the probability of how certain types of interactions (or component states) follow others at a rate that is significantly above chance level: an if–then probabilistic logical function. As a result, the collaborative process can be examined as an evolving, multi-state, Boolean network, and the greater the number of significant transitions between the component states, the greater the complexity of how its evolution unfolds (Kauffman, 1995).
The aforementioned differences in the interactional sequences of WS and IS groups are further illustrated through a qualitative examination of the interactional data. Two pairs of contrasting excerpts are presented to illustrate differences between WS and IS groups. For example, consider the following excerpt,\(^7\) seven utterances long, from the discussion of a group analyzing the first well-structured problem. The three group members are 0201, 0202, 0203 and the prefix “wsp” stands for well-structured problem. The rightmost column lists the codes corresponding to the interactional units within each utterance.

1. **wsp0201** > in the case given to us a man is being fined for speeding we as lawyers have to defend him
   - PA, PA
2. **wsp0202** > we have to defend a man who’s been fined for speeding
   - PA
3. **wsp0203** > ya
   - PC
4. **wsp0201** > the reaction time being 0.8 s the man applied the brakes when the boy had started crossing the road, right?
   - PA, PA
5. **wsp0203** > and the skid length is given
   - PA
6. **wsp0202** > mass of car and man also
   - PA, PA
7. **wsp0201** > friction mu also
   - PA

Within the first three utterances of this excerpt, the group members were able to extract and agree on the problem’s overarching goal (defending a man fined for speeding). They then proceeded with an analysis of the problem, focusing on the relevant parameters (such as the reaction time, skid length, mass, frictional coefficient). What is also evident from this excerpt is that the clearer problem definition afforded by well-structured problems seemed to engender sustained problem analysis activity without much need for critique of that analysis.

In contrast, consider the following excerpt, ten utterances long, from the discussion of a group analyzing the ill-structured version of the same problem. The three group members are 0401, 0402, 0403 and the prefix “isp” stands for ill-structured problem.

1. **isp0401** > while going through the previous record MR. GUPTA has been fined twice he might be in practice of rash driving
   - PA, PA
2. **isp0402** > please wait one should see each & every aspect
   - PC
3. **isp0403** > eye witness also meant to say that there was no fault of the driver.
   - PC
4. **isp0401** > but what petitions we should give in order to defend him
   - PC

---

\(^7\) All excerpts presented here have undergone some minimal editing to make them more readable. In particular, spelling mistakes have been corrected, and unclear short and abbreviated forms have been expanded.
In this excerpt, 0401 attempts to analyze the problem and suggests a possible cause (rash driving). This attempt is immediately met with critique from 0402 and 0403 in utterances 2 and 3 respectively. In utterance 4, 0401 responds by raising another question to counter the preceding critique, thereby furthering the critiquing activity. A similar pattern seems to be repeated in this excerpt, that is, each time an attempt is made at analyzing the problem by suggesting a possible cause (such as the driver’s record, his age, the boy’s fault), the discussion goes into a sustained critique of the preceding analysis. Contrasting this with the previous excerpt, what becomes evident in this one is that the ambiguity in problem definition in ill-structured problems seemed to engender many attempts at problem analysis, each being followed up with a sustained critique of that analysis.

Next, consider another comparison between WS and IS groups. For example, consider the following excerpt, six utterances long, from the discussion of a group developing a solution for the first well-structured problem. The three group members are 0201, 0202, 0203.

1. **wsp0201**: I found the speed as 12.94 m/s
2. **wsp0202**: how did you get the speed?
3. **wsp0201**: use 3rd equation of motion
4. **wsp0202**: yes my answer is matching that of 0201
5. **wsp0203**: so the man can’t be fined
6. **wsp0202**: thats correct

In this excerpt, 0201 develops a solution by calculating the car’s speed and concluding that the driver was not speeding. This is immediately followed by an evaluation of the preceding solution development (i.e., in utterance 2, 0202 questions 0201 how the speed was calculated). Student 0201 responds by elaborating on his or her solution (recall that solution elaboration is part of solution development). This is again evaluated by 0202 who finds that it is consistent with his or her calculations. Finally, 0203 presents the developed solution for acceptance, which the solution receives. More importantly, this excerpt exemplifies the pattern of interactional activity more likely in well-structured problem-solving groups: solution development activity followed by (un-sustained) solution evaluation.
In contrast, consider the following excerpt, 25 utterances long, from the discussion of a group developing a solution for the ill-structured version of the same problem. The three group members are 1501, 1502, 1503.

1. *isp1501* > so i will fight for gupta  
2. *isp1502* > me too but get accurate information  
3. *isp1502* > like how can we get retardation  
4. *isp1503* > suppose  
5. *isp1502* > don’t just suppose  
6. *isp1503* > from the eye witness account we know that it was boys fault  
7. *isp1503* > what do u think  
8. *isp1501* > i agree with 1503  
9. *isp1501* > so mr gupta should pay half the fine as he did not maintain his car  
10. *isp1503* > but this is not mentioned that he didn’t maintain his car properly  
11. *isp1503* > mechanics doesn’t say so  
12. *isp1501* > he does it is mentioned in mechanics column  
13. *isp1502* > but it’s true that his speed was more  
14. *isp1503* > read it again  
15. *isp1503* > 1502 how can u say that  
16. *isp1501* > may be his speed was less than 55km  
17. *isp1503* > we actually dont know  
18. *isp1501* > u r right  
19. *isp1503* > ya thats what I’m trying to tell  
20. *isp1502* > but in this case we can’t find the accurate speed  
21. *isp1503* > u r right 02  
22. *isp1501* > ok so what should we conclude  
23. *isp1503* > according to eye witness it was boys fault completely  
24. *isp1501* > ya it was boys fault  
25. *isp1503* > and traffic conditions were normal

In this excerpt, 1501 evaluates the preceding discussion (not shown in this excerpt) and declares his or her decision. 1502 evaluates 1501’s decision favorably but asks for more accurate information, sending the discussion into even more evaluation. In utterance 6, 1503 draws the focus back to problem analysis to point out that it was the boy’s fault, which is then critiqued. A similar interactional pattern is then seen subsequently, in utterances 9 through 14, and again in utterances 15 through 25; the pattern being: sustained solution evaluation feeding back into problem analysis and critique.

These two excerpts reveal that the greater multiplicity of the solution paths, solutions, and criteria for evaluating solutions afforded by ill-structured problems
resulted in characteristically different and more complex interactional sequences, especially in the form of feedback loops from solution evaluation to problem analysis and critique. In contrast, interactional sequences in the discussions of WS groups were comparatively simpler and exhibited solution development followed by (un-sustained) evaluation.

Group-Level Analysis 3: Convergence in Group Discussions

Convergence was designed as a second measure for understanding emergent patterns in group discussions. A second measure for complexity was necessary for two reasons: first, it provided an additional lens to characterize interactional complexity in group discussions, and second, it provided a means of establishing convergent validity with the first measure presented earlier (Chatterji, 2003). This is particularly important in light of the fact that the two measures of interactional complexity proposed in this study are new, and therefore, at the very least, there is a need for establishing convergent validity between them.

Convergence is a measure of how group members interact and develop a shared understanding of the problem, select a strategy, develop a solution, and manage the problem-solving process (Fischer & Mandl, 2005; Roschelle, 1996). The idea here was simple: the lower the complexity in the interactional dynamics, the higher will be the likelihood of the emergence of convergence therein, and vice versa. Thus, convergence in group discussion was modeled as an emergent property (Jacobson & Wilensky, 2006) of the interactions between group members using methods developed by Kapur, Voiklis, and Kinzer (2007). Kapur et al. (2007) leveraged the complexity theory concepts of emergent simplicity and emergent complexity (Bar-Yam, 2003) to hypothesize a set of theoretically sound yet simple rules: interactions between group members were conceptualized as goal-seeking adaptations that impact the group by either helping it move towards (impact = 1) or away (impact = -1) from its goal, or maintain its status quo (impact = 0). The impact coding scheme facilitated an operationalization of convergence as a Markov walk (Ross, 1996); the mean distance of the Markov walk forming the measure of convergence used in this study (for a fuller description of the model, see Kapur et al., 2007).

Figure 2 suggests that convergence in the discussions of WS groups, $M = .255$, $SD = .281$, was greater than in those of IS groups, $M = .041$, $SD = .337$. Controlling for the effects of group prior knowledge and group English proficiency, an ANCOVA revealed that there was a significant difference, on average, between the convergence of discussions of WS and IS groups, $F = 10.008$, $p = .002$, partial $\eta^2 = .092$, power = .88.

The analyses thus far suggest that not only did IS group discussions exhibit significantly greater complexity in terms of their interactional sequences but that their discussions were significantly more divergent than those of WS groups.
FIGURE 2 Convergence of group discussion across problem type.  

Taken together, therefore, the two proposed measures of complexity did in fact demonstrate convergent validity. Having established this, the next step was to relate the complexity in interactional dynamics with group performance. Just because IS and WS groups differed in terms of their interactional complexity does not automatically imply that this difference, in turn, necessarily explained any observed variation in group performance. One could very well make this assumption but it is not a logical necessity; the link has to be established and this is precisely what follows next.

Group-Level Analysis 4: Convergence and Group Performance

The measure of group performance was operationalized as the quality of solution produced by the group. This was initially problematic because there were no objectively right or wrong answers to the problem scenarios. However, in consultation with the teacher experts, the strategy adopted was to focus on the extent to which groups were able to support their decisions through a synthesis of both qualitative and quantitative arguments, and supporting them with justifiable assumptions. The extent to which groups were able to accomplish this was rated on a scale from 0 to 4 points in units of 0.5 using a holistic rubric shown in Table 4.

Linear regression showed that convergence in group discussion was a significant predictor, $t = 12.253$, $p < .001$, of group performance as evidenced by the quality of group solutions. As a result, the quality of solution produced by WS groups, $M = 2.179$, $SD = 1.317$, was on average better than that of IS groups, $M = 1.420$, $SD = 1.085$ (see Figure 3). Controlling for the effects of group prior knowledge, an ANCOVA revealed that there was a significant difference between
TABLE 4
Rubric for Coding Quality of Group Solution

<table>
<thead>
<tr>
<th>Quality</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Solution weakly supported, if at all</td>
</tr>
<tr>
<td>1</td>
<td>Solution supported in a limited way relying either on a purely quantitative or a qualitative argument with little, if any, discussion and justification of the assumptions made</td>
</tr>
<tr>
<td>2</td>
<td>Solution is only partially supported by a mix of both qualitative and quantitative arguments; assumptions made are not mentioned, adequately discussed, or justified to support the decision</td>
</tr>
<tr>
<td>3</td>
<td>Solution synthesizes both qualitative and quantitative arguments; assumptions made are adequately discussed and justified to support the decision</td>
</tr>
<tr>
<td>4</td>
<td>Solution synthesizes both qualitative and quantitative arguments; assumptions made are adequately discussed and justified to support the decision</td>
</tr>
</tbody>
</table>

*Note. Mid-point scores of .5, 1.5, 2.5, and 3.5 were assigned when the quality of solution was assessed to be between the major units 0, 1, 2, 3, and 4.*

**FIGURE 3** Group performance across problem type.

The performances of WS and IS groups, $F = 7.200$, $p = .009$, partial $\eta^2 = .067$, power = .76.

**Discussion of Group-Level Analyses**

Differences between groups on the various process and outcome measures can easily be explained in terms of the affordances of well- versus ill-structured problems. Because ill-structured problems do not provide a clear problem definition,
IS groups spent proportionally greater amounts of interactional activity on problem analysis, problem critique, and criteria for developing a solution. LSA further revealed that this lack of clarity in problem definition also resulted in sustained critiquing of problem analysis attempts. The larger solution space afforded by ill-structured problems resulted in sustained evaluation of attempts at solution development, which, in turn, fed back into problem analysis and critique. IS groups were less likely to settle on a suggested solution and develop it through sustained activity. Instead, given a larger set of criteria for solution evaluation afforded by ill-structured problems, ideas for solutions were subjected to sustained solution evaluation, which, in turn, fed back into problem analysis and critique. Thus, the discussions of IS groups were, on average, more complex and chaotic, as indicated by the greater numbers and variety of transitions and feedback loops. LSA analyses of interactional sequences as well as subsequent qualitative analysis speak to this. Indeed, as one of the raters put it succinctly during one of the post-coding reliability discussion sessions: “These guys (referring to IS groups) seem to be all over the place!” It is not surprising then that the IS groups found it difficult to converge on the causes of the problem, set appropriate criteria for a solution, and develop a solution. On the whole, therefore, their discussion were not only more complex and chaotic but also divergent. This lack of convergence in group discussion drove down group performance. WS groups, on the other hand, solved problems that offered more defined problem and solution spaces. Thus, their discussions were, on average, more coherent, less complex, and less likely to exhibit complex transitions or feedback loops. As a result, these groups found it relatively easier to converge on the causes of the problem, set appropriate criteria, and develop a solution, which, in turn, resulted in relatively better group performance. Thus, on many counts, IS groups failed compared to WS groups.

Group-to-Individual Transfer: Hierarchical Linear Modeling (HLM)

Given the nested structure of the data with students nested within groups within experimental condition, HLM was carried out. HLM was deemed to be the more appropriate analyses because the post-test individual performances were not independent observations (even though the post-tests were taken individually) simply due to the collaboration that preceded it (Raudenbush & Bryk, 2002). A lack of independence of observations, therefore, ruled out the simpler and commonly used statistical procedures such as ANCOVAs and MANCOVAs because they assume independence of observations (Stevens, 2002). If one were to ignore this assumption and carry out ANCOVAs and MANCOVAs on a nested data set such as this one, it would have resulted in biased parameter estimates and consequently, results that turn out to be significant when in fact they may not be (Raudenbush & Bryk, 2002). HLM corrects for the dependencies to give more reliable and unbiased parameter estimates. Note that, although HLM
is technically a more advanced technique, conceptually, it is still a regression model.

Two hierarchical models were gradually built corresponding to performance on the well- and ill-structured problems post-tests. The measure of performance on the well- and ill-structured problems post-tests was operationalized as the quality of solutions produced by the student using the same holistic rubric that was used for rating the quality of group solutions presented in Table 4. In addition to controlling for confounding factors (e.g., school, problem order), several individual- and group-level predictors were entered, one-by-one. Single degree-of-freedom log-likelihood tests determined the significance of each predictor in the model.

**Performance on Well-Structured Problems Post-Test**

For modeling the performance on the well-structured problems on post-test 1, individual-level variables included prior knowledge (pre-test score) and the level of participation in group discussion (operationalized as the proportion of all interaction units contributed by a group member). Group-level predictors included problem type (well- vs. ill-structured), group prior knowledge, group solution quality, proportion of group activity that was problem centered (operationalized as the sum of proportions PA, PC, CD, SD, and SE activity), and level of participation inequity in group discussions (operationalized as the standard deviation of participation ratios of group members; the higher the standard deviation, the more inequitable the group discussion). The choice of individual and group-level predictors was grounded in previous CSCL research. For example, the role of individual and group prior knowledge, individual participation and level of inequity in group discussions, problem-centeredness of interactional activity, and so on are all important variables that need to be taken into account in explaining the variation in group as well as the eventual individual performance (Barron, 2003; Cohen, Lotan, Abram, Scarloss, & Schultz, 2002; Jonassen & Kwon, 2001; Schellens et al., 2005; Spada, Meier, Rummel, & Hauser, 2005; Suthers, 2006). Individual performance on the well-structured problems post-test 1 formed the outcome variable. Thus, the hierarchical model controlled for a number of relevant individual- and group-level variables, none of which had significant effects on the outcome variable (see Appendix C, Table 6). Controlling for these effects, this analysis suggested that, on average, participants from IS groups performed significantly better on the well-structured problems post-test 1 than their counterparts from the WS groups, \( \chi^2 = 22.82, \ p < .001 \). Of the overall variation in the performance on the well-structured problems, 64% was due to differences between individual participants (between-individual variance) and 36% was due to differences between groups (between-group variance). The experimental condition (IS vs. WS) explained approximately half (53%) of the between-group variance.
Performance on Ill-Structured Problems Post-Test

Individual performance on the ill-structured problems post-test 2 was also modeled as a level 1 outcome variable. All the predictors in the aforementioned model were also included in this one. The only addition was to include individual performance on the well-structured problems post-test 1 as a predictor of performance on the ill-structured problems post-test 2—the outcome variable in this model. This was reasonable because the well-structured problems post-test 1 temporally preceded the ill-structured problems post-test 2 and thus some variation on the latter could possibly be explained by the former. Further, the inclusion was also necessary to test the contrasting-case design (i.e., if indeed solving ill-structured problems followed by well-structured problems should help students perform better on an ill-structured problem subsequently, then performance on the well-structured problems post-test 1 should have a significant effect on that on the ill-structured problems post-test 2). It was precisely this effect that was modeled.

Again, as with the first model, the hierarchical model controlled for a number of relevant individual- and group-level variables, some of which had significant effects on the outcome variable (see Appendix C, Table 7). These effects are interesting findings in and of themselves. In fact, even the non-significance of individual and group prior knowledge on the ill-structured problems post-test performance is an interesting finding. However, for the purposes of this article, these effects, significant or otherwise, form control measures. This allows us to focus on the main effect of problem type: controlling for the effects of a number of relevant individual- and group-level predictors, participants from IS groups outperformed their counterparts from the WS groups on the ill-structured problems post-test, $\chi^2 = 27.21, p < .001$. Importantly, individual performance on well-structured problems post-test 1 was a significant predictor of performance on ill-structured problems post-test 2, $\chi^2 = 30.76, p < .001$. Of the overall variation in the performance on the ill-structured problems, 69% was due to differences between individual participants (between-individual variance) and 31% was due to differences between groups (between-group variance). The experimental condition (IS vs. WS) explained almost all (98%) of the between-group variance.

In sum, results from the two hierarchical models suggested that participants from the IS groups outperformed those from the WS groups on not only the well-structured problems post-test but also the subsequent ill-structured problems post-test. The more counterintuitive and intriguing of the two findings is that students from IS groups outperformed their counterparts from WS groups on the well-structured problems post-test. One would have expected students from the WS groups to be at least as good if not better in solving well-structured problems. After all, they had just solved two such problems in groups before attempting similar ones individually whereas the IS groups had not even solved
such problems. To explain this intriguing finding, content analysis of students’ solutions for the well-structured problems post-test was carried out.

Content Analysis of Students’ Solutions

Content analysis was necessary because the holistic rubric presented in Table 4 provides only an overall rating of solution quality and not finer-grained insights into the content of solutions produced by students; content that may, in turn, provide partial insights into the particular kinds of knowledge and skills that might have been transferred from the collaborative phase to the individual post-test performance. Thus, content analysis of students’ solutions on the well-structured problems post-test was carried out to provide some explanatory support for the statistical findings from the hierarchical models.

Content analysis suggested that three critical elements—problem representation, nature of assumptions, and contextual factors—distinguished the solutions produced by students from IS groups from those produced by students from WS groups. In other words, the frequency of occurrence of these elements in students’ solutions and the manner in which they were combined to develop quantitative and qualitative arguments (which the rubric in Table 4 rated as part of its holistic rating of solution quality) seemed to explain, in part, the better performance of students from IS groups on the well-structured problems post-test. What follows is a content analysis of students’ solutions of one of the two well-structured problem scenarios (see Appendix B2) on the well-structured problems post-test to exemplify the nature of the three elements.

**Problem Representation**

Table 5 provides a comparison between students from WS and IS groups in terms of the typical symbolic representations of the well-structured problem scenario on the well-structured problems post-test.

Table 5 suggests that students from both conditions were fairly similar in identifying the relevant parameters such as initial and final velocities, coefficient of friction, stopping distance, the laws of kinematics, and so on. This was not surprising given that the problem scenario was a well-structured one. However, there were important differences between students from the two conditions in terms of how they represented the identified parameters. These differences are important because how one represents a problem influences the nature of the solution. For example, content analysis of the solutions revealed that 66% of students from WS groups represented the coefficient of friction as a fixed value equal to .7 (the midpoint of the range given in the problem scenario). In contrast, 72% of students from IS groups represented the coefficient of friction as a range from .6 to .8. Similarly, 69% of students from WS groups represented the stopping distance as
TABLE 5
Typical Symbolic Problem Representations of Students from WS and IS Groups

<table>
<thead>
<tr>
<th>Parameters/Laws</th>
<th>Students from WS Groups</th>
<th>Students from IS Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Initial velocity, (u)</td>
<td>Unknown; to be found</td>
<td>Unknown; to be found</td>
</tr>
<tr>
<td>2. Final velocity, (v)</td>
<td>(v = 0)</td>
<td>(v = 0)</td>
</tr>
<tr>
<td>3. Gravitational acceleration, (g)</td>
<td>(g = 9.81 \text{ or } 10)</td>
<td>(g = 9.81 \text{ and } 10)</td>
</tr>
<tr>
<td>4. Coefficient of friction, (\mu)</td>
<td>(\mu = .7)</td>
<td>(.6 \leq \mu \leq .8)</td>
</tr>
<tr>
<td>5. Acceleration, (a)</td>
<td>Uniform, (a = -\mu g)</td>
<td>Non-uniform, (a = -\mu g)</td>
</tr>
<tr>
<td>6. Stopping distance, (s)</td>
<td>(s = 11 \text{ m})</td>
<td>(s \geq 11 \text{ m})</td>
</tr>
<tr>
<td>7. Reaction time, (t)</td>
<td>(t = .8 \text{ s})</td>
<td>(.7 \leq t \leq .9)</td>
</tr>
<tr>
<td>8. Mass of car + driver, (m)</td>
<td>(m = 875 \text{ kg})</td>
<td>(m = 875 \text{ kg})</td>
</tr>
<tr>
<td>9. Laws of kinematics</td>
<td>(v^2 = u^2 + 2as)</td>
<td>(v^2 = u^2 + 2as)</td>
</tr>
<tr>
<td></td>
<td>(F = \mu mg)</td>
<td>(F = \mu mg)</td>
</tr>
</tbody>
</table>

Representing the coefficient of friction, stopping distance, reaction time, and gravitational acceleration as fixed values, as opposed to a range of values, is critical because the latter is not only a better representation of the problem but also requires a commitment to exploring the sensitivity of prospective solutions to the range of variation of the respective parameters. This was particularly relevant because the problem scenarios on the well-structured problems post-test were designed to be sensitive to parametric variation in the values of the coefficient of friction and stopping distance but not to reaction time and gravitational acceleration. Hence, testing the sensitivity of solutions to the variation in parameter values was an important expert-like skill that differentiated the solutions produced by students from IS from those of students from WS groups. Whereas 59% of students from IS groups attempted some form of sensitivity analysis in their solutions, the same was true for only 23% of students from the WS groups.

**Nature of Assumptions**

An important concept in Newtonian kinematics deals with the understanding of the conditions under which the laws of kinematics can be applied. This set of conditions is usually referred to as “ideal conditions,” wherein the motion can be assumed to be rectilinear, objects can be assumed to perfectly rigid, acceleration
can be assumed to be uniform, and so on; conditions that usually cannot be assumed under real situations as presented in the problem scenarios on the WSP post-test.

Content analysis of solutions revealed that 89% of students from WS groups assumed ideal conditions whereas only 33% of students from IS groups did the same. For example, in addition to assuming a fixed value for the coefficient of friction and gravitational acceleration, students from WS groups also tended to assume that deceleration from the time the brakes are applied to the time the car comes to a complete stop is uniform. Assuming uniform deceleration allowed them to apply the laws of kinematics. In contrast, students from IS groups tended to model the deceleration period as a piece-wise function (i.e., the car decelerates at slower rate first up until the point it starts to skid, after which it decelerates at a faster rate).

**Contextual Parameters**

Another difference between the content of the solutions produced by students from the two conditions was the extent to which students put forth qualitative arguments arising from contextual parameters that could possibly mediate the nature and support for their decisions. For example, content analysis revealed two contextual parameters to be particularly important: (a) the car being brand new and (b) the road being an inner-city road.

With regard to the car being brand new, 81% of students from IS groups, as opposed to 32% of students from WS groups, picked up on this given information to qualitatively argue that the condition of the braking system and traction provided by the tires were likely very good. Assuming the condition of the braking system and the traction to be good, in turn, was used to argue that the stopping distance was probably not too much greater than the length of the skid marks and also that the coefficient of friction was likely on the higher end of the range given in the car’s information manual. As for the road being an inner city road, 75% of students from IS groups, as opposed 17% of students from WS groups, used this information to argue that the road may not have been well-maintained (a reasonable argument if one is familiar with inner-city roads in the city in question). Consequently, a qualitative argument was made for the coefficient of friction on an inner city road to be on the higher end of the range given in the car’s information manual.

Thus, content analysis suggested that students from IS groups were more likely to use the two contextual parameters to constrain the range of values of the coefficient of friction, which helped provide qualitative support for the quantitative sensitivity analysis mentioned earlier. In fact, at a conceptual level, using the aforementioned contextual parameters to constrain the value of the coefficient of friction also reflected a better understanding of the very concept of the coefficient of friction (i.e., the coefficient of friction is not a property of a surface in isolation but a joint property of the surfaces in contact [the tires and the road], and therefore,
factors that may indirectly affect the roughness of the surfaces in contact ought to be taken into consideration).

**Discussion of Content Analysis**

The content analysis provided finer-grained insights into the nature of the three elements—problem representation, nature of assumptions, and contextual parameters—distinguished, in part, the solutions produced by students from IS groups from those produced by students from WS groups. However, it was not the case that all IS group students exhibited the aforementioned elements in their solutions. Nor was it the case that none of the WS group students exhibited the aforementioned elements in their solutions. It was the combination of these elements and the frequency of their occurrence in the solutions produced by IS group students that resulted in their solutions being of a better quality on average. This is because the holistic rubric in Table 4 assigned higher ratings when these elements were present because these elements were evidence of better quantitative and qualitative arguments supporting their decisions. Hence, it is important that the content analysis not be interpreted in a reductive manner simply because the different elements are described separately.

It is also important to note that the elements, in addition to contributing to the overall quality of solutions, also functioned to structure the problem scenario. For example, it became evident from the students’ solutions that the nature of problem representation structured the problem and solution spaces (Goel & Pirolli, 1992; Voss, 2005). For example, representing the coefficient of friction as a range of values structured the problem space in a way that allowed for the possibility of testing the sensitivity of the solution space to parametric variation in the value of the coefficient of friction. In contrast, representing the coefficient of friction as a fixed value also structured the problem and solution spaces but in a (different) way that did not allow for testing the sensitivity of the solution space to parametric variation in the value of the coefficient of friction. Thus, the better the representation, the better was the structuring of the problem scenario (Chi, Feltovich, & Glaser, 1981). Following a similar line of reasoning, it is easy to see how the nature of assumptions and qualitative reasoning with contextual parameters also functioned to structure the problem scenario (Goel & Pirolli, 1992).

**GENERAL DISCUSSION**

This study experimentally manipulated the level of structure in problematic tasks given to collaborative groups in CSCL settings. As hypothesized, IS groups engaged in discussions that were, on average, more complex, chaotic, and divergent when compared to those of WS groups. This adversely affected the group performance of IS groups; WS groups outperformed IS groups on the quality of solutions.
they produced. However, the study also demonstrated that the contrast between ill-structured problems followed by well-structured problems facilitated a spontaneous transfer of problem-solving skills, which, in the absence of the contrast, might have remained unrealized. The data provided strong evidence for this: first, participants from IS groups performed significantly better on the ill-structured problems post-test compared to their counterparts from the WS groups; and second, their performance on the well-structured problems post-test was a significant predictor of their subsequent performance on the ill-structured problems post-test. Therefore, despite the greater struggle, complexity, and divergence in the discussions of IS groups resulting in failure, participants from IS groups outperformed those in WS groups on both well-structured and ill-structured problems post-tests, in turn, demonstrating an existence proof for productive failure. Of course, this opens the door to the argument that the rich, open-ended, and complex nature of collective discourse in IS groups was in itself a structure for students in these groups. Indeed, this is precisely what one reviewer of this manuscript pointed out in terms of what is meant by productive failure, and it only bolsters the argument for allowing such structures to emerge from within as opposed to imposing them from the outside.

Although the findings evidently suggest that the contrasting-case design of solving ill-structured problems followed by well-structured problems facilitated a spontaneous transfer of problem-solving skills for IS group students, they do not quite explain how or what was transferred during the contrast. Content analysis of students’ solutions on the well-structured problem post-test (which provided the contrast) could well provide some insights into what may have been transferred. Concomitantly, the argument from discernability, as hypothesized, could offer an explanation for how the transfer may have been facilitated during the contrast. Indeed, it must be acknowledged that without a more in-depth analysis it is hard to ascertain exactly what was transferred. Notwithstanding, an attempt at a plausible explanation can still be made.

Recall that the IS group students received a problem sequence in which the level of problem structuredness was contrasted (ill- followed by well-structured problems) while keeping the target content in Newtonian kinematics constant. Findings from the group-level analyses suggested that IS groups struggled to deal with the complexity of the ill-structured problems. This was evident in their interactional activity centered on problem analysis and criteria development, as well as sustained problem critique and solution evaluation with a number of transitions and feedback loops. In other words, although seemingly unproductive, a more complex and divergent exploration of the problem and solution spaces was what differentiated the interactional dynamics of IS groups from those of WS groups. From a functional standpoint, this complex and divergent exploration of the problem space through repeated attempts at analyzing the problem, defining criteria, and developing possible solutions can be seen as attempts, albeit unsuccessful,
at structuring the problem and solution spaces even though they were met with sustained critique and evaluation and did not lead to productive outcomes in the short term (Goel & Pirolli, 1992; Spiro, Coulson, Feltovich, & Anderson, 1988; Sweller, 1988).

After solving ill-structured problems in groups, when IS group students were asked to solve well-structured problems individually, they outperformed their counterparts from the WS groups. This finding is counterintuitive but consistent with the argument from discernability that keeping the target content the same and varying the level of problem structuredness may help students from the IS groups discern how to structure an ill-structured problem (Marton, 2006). Perhaps students from the IS groups discerned how the well-structured problems represented ways in which ill-structured problems could be structured. Content analysis of students’ solutions on the well-structured problems post-test provides some support for this explanation as it suggests three elements—problem representation, nature of assumptions, and contextual parameters—that distinguished the solutions produced by students from IS groups from those produced by students from WS groups. IS group students produced better quantitative representations of targeted content, assumptions, and qualitative reasoning with contextual parameters to structure the problem and solution space of the well-structured problems. Indeed, these were the elements that resulted in students from IS groups outperforming their counterparts from WS groups on the well-structured problems. Furthermore, it seems that these elements for structuring a problem scenario were learnt during the contrast provided by well-structured problems and not in the preceding collaborative problem-solving of ill-structured problems. That said, solving ill-structured problems first perhaps better prepared students from the IS groups to learn from their problem-solving efforts on the well-structured problems post-test (Schwartz & Bransford, 1998).

Subsequently, when solving the ill-structured problems individually, students from the IS groups again outperformed their counterparts from the WS groups. A plausible but partial explanation for this finding could be that students from the IS groups were better able to structure the ill-structured problem scenarios by using quantitative representations of the targeted content, assumptions, and qualitative reasoning with contextual parameters. Needless to say, an in-depth content analysis of students’ solutions on the ill-structured problems post-test is needed for a more direct and fuller explanation. However, note that the solutions on the ill-structured problems post-test were rated for overall quality using the holistic rubric in Table 4. As described earlier, the rubric assigned higher ratings to solutions in which the three elements were present and combined to provide quantitative and qualitative support for the solutions than when they were not. This provides indirect evidence that the solutions to the ill-structured problems produced by IS group students contained these elements for without them, the overall quality ratings for solutions produced by students from IS groups would
not have been significantly better than those produced by students from WS group.

Thus, the explanation that perhaps what students in IS groups learned was how to structure an ill-structured problem seems plausible. Solving ill-structured problems influenced how they solved and learned from well-structured problems, which, in turn, helped them discern how to structure ill-structured problems (Marton, 2006; Schwartz & Bransford, 1998).

Limitations

The scope of inference of any observed effect technically holds only under the conditions and settings of the respective study. Thus, findings from this study may not be extended to other communication modalities, age-groups, and cultures. Although special care was taken in randomly grouping and assigning participants to the different treatments from within this sample, the population from which the sample is drawn is constrained and, therefore, the sample may still be considered a sample of convenience. However, one must note that this study is one of the larger-scale studies undertaken in CSCL research, involving seven educational institutions and more than three hundred students. This bears favorably on the study’s external validity.

It is also possible that the narrow bandwidth afforded by chat communication may have differential effects for the IS and WS groups thereby confounding the study’s results and findings. Although this effect can never be fully mitigated, careful design and validation of the study’s instruments did help alleviate this concern. Pilot studies with the previous cohort of eleventh-grade students and teachers helped refine the problem scenarios and the study’s design to mitigate this limitation as far as possible. This process has been described in detail. The pilot studies also did not reveal or suggest any such differential effects. Plus, the fact that the chat environment used for this study was one that students used on a daily basis for collaborating and chatting also worked to the study’s advantage. Notwithstanding, extending this study’s findings to other communication modalities (face-to-face or with other CSCL tools) will only help unpack this effect further.

Another limit on the scope of inference comes from the nature of the target domain on which the learning outcomes were examined in this study. The target domain of Newtonian kinematics is itself a fairly well-structured domain. It consists of a finite set of parameters (e.g., mass, displacement, velocity, acceleration, time) interacting with each other in a pre-defined and predictive manner (e.g., the three equations, laws of motion). In other words, this target domain consists of processes that are ontologically direct. Hence, it is problematic to extend the findings from this study to target domains that are not as well-structured, for example, target domains that consist of processes that are ontologically emergent. For instance, a target domain such as Genetic Networks involves
PRODUCTIVE FAILURE

multiple components interacting with each other in a number of direct and indirect ways. From these interactions, complex phenotypic patterns emerge. Hence, conceptions of processes in Newtonian kinematics may not be applicable to target domains that are ontologically different from it, such as Genetic Networks. Indeed, Chi (2005) argues that when conceptions of direct processes are extended to understand emergent processes, they may not only lead to misconceptions of the emergent processes but that these misconceptions tend to be highly resistant to instructional intervention as well. This also explains the recent push in the learning sciences to investigate how learners learn and come to understand complex systems and emergent processes (Jacobson & Wilensky, 2006).

Finally, the two measures of interactional complexity are new and much more research needs to be done to validate them. As with any new set of measures, there will always be more questions than answers. Although the proposed measures were derived from commonly used and fairly robust measures in complex systems science (i.e., Boolean networks and simple rule-based computational models) and do give us some insight into the interactional complexity of group discussions, they are by no means the best possible. In the light of the recent calls in the cognitive sciences (Goldstone & Janssen, 2005) and the learning sciences (Jacobson & Wilensky, 2006) to leverage theory and methods in complexity to better understand learning and problem solving, this study presents an initial effort in that direction. Indeed, this is one area for future research to build on as is articulated in the next section.

Future Research

Suggestions for future research may be broadly categorized into three major areas. First, future research would do well to examine the productive failure hypothesis using other definitions and forms of structure on the learning and performance space of learners. For example, research could keep the level of structure in the problems the same while varying the structure in the problem-solving process through the provision of process scaffolds, argumentation tools, representational tools, and so on. Carrying out a comparative study similar to the one reported herein but with a different form of structuring of the learning and performance space forms a natural and immediate extension. Concomitantly, extending this work to target domains other than Newtonian kinematics is also important. Such replications and extensions of productive failure research with different forms of structuring form the focus of our current research program, specifically in the domain areas of mathematics and biology.

Second, what also needs immediate analysis is the variation within the IS groups. Although, on average, IS groups exhibited productive failure, it was also clear that some failed more than others in the shorter term inasmuch as some gained more than others in the longer term. What explains this variation within
the IS groups? Examining the nature of interactional behaviors and relating them to eventual gains in group and individual performance would be most insightful. For example, preliminary analysis suggests individual, behavioral characteristics (as manifested in the interactions) of *persistence, tenacity, inventiveness, and persuasiveness* to be critical. Indeed, as reviewers insightfully pointed out, examining productive failure from cognitive aspects alone is not sufficient; learners’ frustration thresholds and level of engagement in solving the problem, for example, may be particularly critical as well; low frustration thresholds and engagement levels may not result in productive failure. The aforementioned interactional behavioral characteristics, particularly persistence and tenacity, speak to these concerns and variation on these preliminarily seems to add further explanatory power to productive failure. However, such micro-genetic analysis is only in its infancy and much more work, including follow-up studies, has to be carried out before any meaningful findings emerge.

Finally, such micro-genetic interactional analysis also has the potential of feeding into computational models of collective behavior (Epstein & Axtell, 1996; Goldstone, 2006). This study demonstrated how conceptualizing group behavior (e.g., interactional sequences, convergence in group discussion) as an emergent property of the local interactions among group members provides insight into collaborative processes and outcomes. A way forward is to further leverage the theory of and methods from complex systems science to provide additional insights into the interactional dynamics of collaboration (Jacobson & Wilensky, 2006; Goldstone, 2006). Micro-genetic interactional analysis can be used to define and build agent-based models of collective behavior, findings from which can inform further analysis and experiments. Iterating back and forth between *in-situ* and *in-silico* experiments (Epstein & Axtell, 1996; Goldstone & Janssen, 2005) may potentially further our understandings of the collective dynamics of problem solving to a level that may not be possible otherwise (Kapur et al., 2007).

**CONCLUSION**

In ending, perhaps it is best that I delineate what this study claims from what it does not. This study was designed to show that, under certain conditions, students’ engagement in solving ill-structured problems—problems that are beyond their skill sets and abilities—can be a productive exercise in failure. The data provide evidence for and support the productive failure hypothesis. However, the claim is not that one should not structure learner experiences at all. Instead, cognizant of the fact that the very act of designing problem scenarios and problem-solving sequences is already an externally imposed structure on the learning and performance space, this study claims that one need not necessarily provide *additional* structures within ill-structured problem-solving activities; the implication
being that, not overly structuring the problem-solving activities of learners, and permitting students to struggle and possibly even fail can be a productive exercise in failure. Of course, believing in the efficacy of structuring what might otherwise be a complex, divergent, and unproductive process is well-placed (Kirschner et al., 2006). However, allowing for the concomitant possibility that, under certain conditions, even ill-structured, complex, divergent, and seemingly unproductive processes have a hidden efficacy about them requires a paradigm shift—in theory and in practice. Resisting the rarely questioned, near-default rush to structure problem-solving activities, perhaps it might be fruitful to first investigate conditions under which ill-structured problem-solving activities lead to productive failure as opposed to just failure.

ACKNOWLEDGMENTS

The research reported in this article was funded in part by the Spencer Research Training Grant and the Education Policy Research Fellowship from Teachers College, Columbia University to the author. I would like to thank the students, teachers, and principals of the participating schools for their support for this project. I am also deeply indebted to Professor Charles Kinzer for his support and guidance throughout the project. Finally, I am grateful to Professors David Hung, Donald J. Cunningham, Florence Sullivan, Liam Rourke, John Voiklis, and the reviewers of this manuscript for their insightful comments and suggestions.

REFERENCES


**APPENDIX A: SAMPLE ITEMS FROM THE 25-ITEM MCQ PRE-TEST**

1. Two cars having different weights are traveling on a level surface with different but constant velocities. Within the same distance, greater force will always be required to stop the car with the greater
   (A) weight  
   (B) velocity  
   (C) kinetic energy  
   (D) momentum

2. A 5 kg block is resting on a rough horizontal plane. The coefficient of friction between the block and the plane is 0.8. A 50 N force parallel to the plane is applied on the block for 10 seconds and then removed. The block eventually comes to a stop. Assuming $g = 10 \text{ ms}^{-2}$ and that the coefficient of friction does not change, the total distance traveled by the block equals
   (A) 20 m  
   (B) 25 m  
   (C) 100 m  
   (D) 125 m

3. A car starts moving from rest in a straight line with a constant acceleration of 5 ms$^{-2}$, then at constant velocity, and finally decelerating at the rate of 5 ms$^{-2}$ before coming to a stop. If the total time of motion equals 5 s and the average speed for the entire motion equals 4 ms$^{-1}$, how long does the car move at constant velocity?
   (A) 1 s  
   (B) 2 s  
   (C) 3 s  
   (D) 4 s
APPENDIX B1: COLLABORATIVE PHASE
PROBLEM SCENARIOS

Ill-Structured Problem 1

You have recently been hired as a lawyer for a prestigious law firm. On your first
day, you are sent to meet with an important client who has been fined for speeding.
Opening your work file, you find your assignment:

Dear new lawyer,
This morning, I received a call from Mr. Gupta asking me for help. According to
him, he almost ran over a small boy this morning in downtown Ghaziabad and was
fined for speeding. He insists that he was not. He says that the boy suddenly ran on
to the road and he braked very hard and managed to avoid an accident. However,
this was enough for a policeman who happened to be there to fine him Rs. 20,000
for speeding. Mr. Gupta is a very important client of our firm and we must do our
best to help him. I trust you will give this case your best effort. I am attaching his
file for your reference.
I am meeting with Mr. Gupta later this evening. So, I need you to investigate
this case and submit your report to me with your analyses and recommendation
by today.
Sincerely,
Nitin Sharma
Senior Partner
PS—Please note that the word of law is very clear on this. A person is speeding if
and only if he is driving above the legal speed limit of the road. No exceptions.

CLIENT FILE

Name: Mr. Amit Gupta
Age: 52 yrs
Driving Experience: 34 yrs
Prior Traffic Violations: 1981 (Fined for speeding, Rs. 500),
1993 (Fined for drunk driving, Rs. 10,000)

To carry out your investigation, you go through a number of steps such as (a)
interviewing an eye-witness, (b) analyzing the incident report filed by traffic
police, (c) accessing the medical examination reports, and (d) interviewing the
mechanic who inspected the car after the incident.

EYE-WITNESS’ ACCOUNT

“I was walking on the roadside pavement. I don’t recall the traffic on the road to be
particularly heavy. Suddenly, I noticed a small boy run out on to the road chasing
a cricket ball. The next thing I heard was a loud screeching sound. I realized that it came from an Ambassador car skidding to a stop in order to avoid running the boy over. The boy was very lucky to have escaped any injury. I think the boy took about 3 seconds to cross the road, but I don’t think he looked at the traffic before crossing the road. He was just chasing the ball!”

**TRAFFIC POLICE INCIDENT REPORT**

- Traffic conditions: *Normal*
- Weather conditions: *Bright and sunny; dry road*
- No evidence of a collision between the car and the boy.
- Number of passengers in the car besides the driver: *None*
- Evidence of skid marks: *about 15 meters*
- Speed limit on the road: 55 kmph
- Width of the road: *about 4.5 meters*

**MEDICAL EXAMINATION REPORT**

**General Comments:**
Neither the driver nor the boy sustained any physical injury whatsoever.

**Results of the car driver’s medical tests**

- BP (Blood Pressure) = 110/80
- HR (Heart Rate) = 80
- Weight = 75 kg
- Reaction Time = 0.8 seconds on an average
- Drug/Alcohol Screen = Negative

**MECHANIC**

**You:** What can you say about the condition of the car from your inspection?

**Mechanic:** Well, this is a heavy car weighing about 1570 kg and I can clearly see some wear and tear of the tires and the braking system. The braking fluid is also running out. As a result, the traction between the tires and the road does not seem to be as good as it can be.
You: Oh! Does this mean the car was not maintained properly?
Mechanic: Not really. You see, the traction also depends on the condition of the road. The coefficient of friction between the car’s tires and the road is usually between 0.6 and 0.7. So, given the city’s roads, the level of traction not being as good is quite understandable.
You: So, what are you saying?
Mechanic: What I’m saying is that although the traction is not as good as it could have been, this is quite normal in Ghaziabad. Also, it is hard to tell how much of the wear and tear happened during the skidding itself.
You: OK. Thank you for your time.

Well-Structured Version of Ill-Structured Problem 1

You are a lawyer in a prestigious law firm. You’ve been assigned the following case:

A man was driving his car when, suddenly, a small boy ran out on to the road chasing a ball. He slammed on the brakes and skidded to a stop, leaving a 15 m long skid mark on the road. Luckily the boy was not hurt, but a policeman watching from the sidewalk walked over and fined the man for speeding. An investigation found out that the speed limit on the road is 55 kmph. It also determined that the coefficient of friction between the tires and the road was 0.6. The man’s mass was 75 kg and his reaction time, on average, was found to be about 0.8 seconds. The car’s information manual indicated the mass of the car to be 1570 kg. Witnesses say that the boy took about 3 seconds to cross the 4.5 m wide road.

As the man’s lawyer, will you fight the fine in court? Present your case as best you can.

APPENDIX B2: EXAMPLES OF WELL- AND ILL-STRUCTURED POST-TEST PROBLEM SCENARIOS

Well-Structured Problem

You are driving your brand new car on a straight inner-city road when you suddenly see a goat at a distance directly in front of your car. You slam on the brakes right away and just about manage to avoid crashing into the goat. A traffic policeman comes right over and immediately gives you a speeding ticket. Still shaking from the experience, you decide to investigate the accident scene to see if you should fight the ticket in court. You look around for the signpost indicating the speed limit and find it to read 50 kmph. Looking at the skid marks left by your car, you estimate
their length to be about 11 meters. You consult with your car’s information manual to find that the car weighs 800 kg and that the coefficient of friction between the car tires and road during braking is usually between 0.6 and 0.8. You estimate your own weight to be 75 kg and reaction time to be about 0.8 seconds. Will you fight the ticket in court? Present your case as best you can.

Ill-Structured Problem

Anita was on her way to school on a beautiful mild day in November. Just as she approached an intersection, she noticed that the light had turned yellow and pedestrians were already starting to move across the intersection. She slammed on her brakes and came to a screeching halt. A large delivery truck was following closely behind Anita’s car. Anita heard brakes screeching behind her and then felt the truck crash into the rear of her car. The impact shoved her car into the intersection before it came to a stop. Luckily, she escaped serious injury, as did the driver of the truck. Not surprisingly, her car was badly damaged. The driver of the delivery truck came over to her and said, “Miss, are you crazy—why did you slam on your brakes? The light was just turning yellow! This accident is your fault!”

Two policemen arrived on the scene to investigate the accident. The delivery truck driver ran over to the policeman and repeated his statements. Anita was still recovering from the incident; her chest hurt from the shoulder strap of her seatbelt, but other than that she felt OK.

The policemen then interviewed some witnesses. Most of them told him that both the car and the truck were speeding and almost ran over some pedestrians who had started to cross the road as the light turned yellow. Others, however, felt that the car and truck were traveling at a normal speed for the area. After taking full statements from the witnesses, Officer Singh went back to the police station, and started preparing to file his report on the accident. He began by reading through the various statements and reports.

**ANITA’S STATEMENT**

I was heading to class in my car, when the light at the intersection turned yellow. I saw students starting to cross and so I hit my brakes, coming to a stop. I heard more screeching of brakes; then a truck crashed into the back of my car, dragging my car into the intersection along with it. I was traveling under the speed limit, but the driver of the truck must have been speeding. Did you get the truck driver’s insurance information?
TRUCK DRIVER’S STATEMENT

I had just delivered some boxes to Delhi Public School. This high school kid in an old beat-up car was in front of me—when all of a sudden, she slams on her brakes. So did I, but I still slammed into her. The light had just turned yellow; we both had time to go through the light. This is her fault!

REPORT SUBMITTED BY THE ACCIDENT ANALYSIS SPECIALIST

At the scene of the accident, there were skid marks showing that both vehicles had braked. The speedometer of the truck was stuck at 24 kmph, leading me to believe that the truck was traveling at least 24 kmph when it crashed into the car.

The car was dragged 9 meters into the intersection, coming to a stop close to the northeast corner of the intersection. The truck was found directly behind the car leading me to believe that it had dragged the car along with it. Both vehicles showed straight skid marks from the south side of the intersection to the position where they came to a stop. My analysis of the measurements shows that the car left a 12-meter skid mark prior to being struck by the truck, while the truck left a 6 meter skid mark before hitting the car.

I estimate the average deceleration on the road to be roughly 6.0 m/s². I also noted that the speed limit on the road is 40 kmph. The truck weighs approximately twice the weight of the car. Since we know that the car had stopped just prior to being hit by the truck, the speed of the truck before impact can be estimated to be about 1.5 times the speed of the two vehicles just after impact.

Question
Should Officer Singh charge either driver for speeding? If so, whom should he charge for the accident? Present your case as best you can.

APPENDIX C: HLM MODELS

HLM Model of Individual Performance on the Well-Structured Problem-Solving Post-Test

The outcome variable was individual performance score on well-structured problem (WSP) post-test. Level 1 explanatory variables included individual prior knowledge (IPK) and individual participation ratio (IPR). Variation between groups was explained by problem type (PT) (ill-structured = 1, well-structured = 0) the problem order (PO), group prior knowledge (GPK), and group performance (GP). The effect of an individual’s participation (participation ratio) during
collaborative problem solving was further explained by participation inequity (PI) in a group as well as the proportion of problem-centered interactional activity (PCA) in a group. This resulted in the following model:

\[
WSP_{ij} = \beta_{0j} + \beta_{1j}(IPK_{ij} - IPK_{..}) + \beta_{2j}(IPR_{ij}) + \varepsilon_{ij}
\]

\[
\beta_{0j} = \gamma_{00} + \gamma_{01}(PT_{j}) + \gamma_{02}(PO_{j}) + \gamma_{03}(GPK_{j} - GPK_{..})
\]

\[
+ \gamma_{04}(G}_{j} - GP_{..}) + u_{0j}
\]

\[
\beta_{1j} = \gamma_{10}
\]

\[
\beta_{2j} = \gamma_{20} + \gamma_{21}(PI_{j} - PI_{..}) + \gamma_{22}(PCA_{j} - PCA_{..})
\]

where subscripts \(i\) and \(j\) refer to levels 1 and 2 respectively. Table 6 presents the model summary with estimation of effects using robust standard errors.

HLM model of Individual Performance on the Ill-Structured Problem-Solving Post-Test

The outcome variable was individual performance score on the ill-structured problem (ISP) post-test. The other individual and group level variables remained the same except performance on WSP post-test was included as a level 1 individual variable. This resulted in the following model:

\[
ISP_{ij} = \beta_{0j} + \beta_{1j}(IPK_{ij} - IPK_{..}) + \beta_{2j}(IPR_{ij}) + \beta_{3j}(WSP_{ij}) + \varepsilon_{ij}
\]

\[
\beta_{0j} = \gamma_{00} + \gamma_{01}(PT_{j}) + \gamma_{02}(PO_{j}) + \gamma_{03}(GPK_{j} - GPK_{..})
\]

\[
+ \gamma_{04}(G}_{j} - GP_{..}) + u_{0j}
\]

\[
\beta_{1j} = \gamma_{10}
\]

\[
\beta_{2j} = \gamma_{20} + \gamma_{21}(PI_{j} - PI_{..}) + \gamma_{22}(PCA_{j} - PCA_{..})
\]

\[
\beta_{3j} = \gamma_{30}
\]

Table 7 presents the model summary with estimation of effects using robust standard errors.
### TABLE 6
Summary of Robust Estimates for the Two-level Model for Individuals’ Well-structured Problem-Solving Performance

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Coefficient</th>
<th>SE</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma_{00}$ : Intercept</td>
<td>2.24</td>
<td>.47</td>
<td>.000*</td>
</tr>
<tr>
<td>$\gamma_{01}$ : Problem Type (PT)</td>
<td>2.02</td>
<td>.33</td>
<td>.000*</td>
</tr>
<tr>
<td>$\gamma_{02}$ : Problem Order (PO)</td>
<td>-.446</td>
<td>.31</td>
<td>.152</td>
</tr>
<tr>
<td>$\gamma_{03}$ : Group Prior Knowledge (GPK)</td>
<td>.104</td>
<td>.06</td>
<td>.410</td>
</tr>
<tr>
<td>$\gamma_{04}$ : Group Performance (GP)</td>
<td>.128</td>
<td>.13</td>
<td>.319</td>
</tr>
<tr>
<td>$\gamma_{10}$ : Individual Prior Knowledge (IPK)</td>
<td>.012</td>
<td>.01</td>
<td>.229</td>
</tr>
<tr>
<td>$\gamma_{20}$ : Individual Participation Ratio (IPR)</td>
<td>1.57</td>
<td>1.17</td>
<td>.185</td>
</tr>
<tr>
<td>$\gamma_{21}$ : Participation Inequity (PI)</td>
<td>-.5.25</td>
<td>6.96</td>
<td>.431</td>
</tr>
<tr>
<td>$\gamma_{22}$ : Problem-centered Activity (PCA)</td>
<td>6.20</td>
<td>4.46</td>
<td>.163</td>
</tr>
</tbody>
</table>

**Random Effect**

<table>
<thead>
<tr>
<th>Variance Component</th>
<th>$\tau^2_0$ : L2 variance</th>
<th>0.94</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma^2$ : L1 variance</td>
<td>4.22</td>
<td></td>
</tr>
</tbody>
</table>

**Deviance**

1246.56

*Significant effect.

Note: Problem Type $= 0$ for well- and 1 for ill-structured problems.

### TABLE 7
Summary of Robust Estimates for the Two-level Model for Individuals’ Ill-structured Problem-Solving Performance

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Coefficient</th>
<th>SE</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma_{00}$ : Intercept</td>
<td>.030</td>
<td>.31</td>
<td>.923</td>
</tr>
<tr>
<td>$\gamma_{01}$ : Problem Type (PT)</td>
<td>1.41</td>
<td>.20</td>
<td>.000*</td>
</tr>
<tr>
<td>$\gamma_{02}$ : Problem Order (PO)</td>
<td>-.642</td>
<td>.19</td>
<td>.001*</td>
</tr>
<tr>
<td>$\gamma_{03}$ : Group Prior Knowledge (GPK)</td>
<td>-.020</td>
<td>.03</td>
<td>.501</td>
</tr>
<tr>
<td>$\gamma_{04}$ : Group Performance (GP)</td>
<td>.244</td>
<td>.07</td>
<td>.001*</td>
</tr>
<tr>
<td>$\gamma_{10}$ : Individual Prior Knowledge (IPK)</td>
<td>.008</td>
<td>.01</td>
<td>.344</td>
</tr>
<tr>
<td>$\gamma_{20}$ : Individual Participation Ratio (IPR)</td>
<td>1.69</td>
<td>.85</td>
<td>.047*</td>
</tr>
<tr>
<td>$\gamma_{21}$ : Participation Inequity (PI)</td>
<td>-6.55</td>
<td>5.80</td>
<td>.260</td>
</tr>
<tr>
<td>$\gamma_{22}$ : Problem-centered Activity (PCA)</td>
<td>1.92</td>
<td>2.66</td>
<td>.472</td>
</tr>
<tr>
<td>$\gamma_{50}$ : Well-structured Problem-solving</td>
<td>.215</td>
<td>.04</td>
<td>.000*</td>
</tr>
</tbody>
</table>

**Random Effect**

<table>
<thead>
<tr>
<th>Variance Component</th>
<th>$\tau^2_0$ : L2 variance</th>
<th>0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma^2$ : L1 variance</td>
<td>2.05</td>
<td></td>
</tr>
</tbody>
</table>

**Deviance**

1011.52

*Significant effect.

Note: Problem Type $= 0$ for well- and 1 for ill-structured problems.