Problem-Based Learning in K–12 Education : Is it Effective and How Does it Achieve its Effects?

Clarice Wirkala and Deanna Kuhn

Am Educ Res J 2011 48: 1157
DOI: 10.3102/0002831211419491

American Educational Research Journal
October 2011, Vol. 48, No. 5, pp. 1157–1186
DOI: 10.3102/0002831211419491
© 2011 AERA. http://aerj.aera.net

Problem-Based Learning in K–12 Education: Is it Effective and How Does it Achieve its Effects?

Clarice Wirkala
Deanna Kuhn
Teachers College Columbia University

Enthusiasm for problem-based learning (PBL) is widespread, yet there exists little rigorous experimental evidence of its effectiveness, especially in K–12 populations. Reported here is a highly controlled experimental study of PBL in a middle school population. Between- and within-subject comparisons are made of students learning the same material under three instructional conditions: lecture/discussion, characteristic small-group PBL, and solitary PBL. Assessments of comprehension and application of concepts in a new context 9 weeks after instruction showed superior mastery in both PBL conditions, relative to the lecture condition, and equivalent performance in the two PBL conditions, the latter indicating that the social component of PBL is not a critical feature of its effectiveness.

KEYWORDS: instruction, problem-based learning, collaboration, learning skills, retention, knowledge

Problem-based learning (PBL) is a teaching and learning method in which students engage a problem without preparatory study and with knowledge insufficient to solve the problem, requiring that they extend existing knowledge and understanding and apply this enhanced understanding to generating a solution. Problems are "ill-structured" ones that do not have a single, clear-cut or formulaic solution, motivating students to ask questions and to seek additional information. This is the core definition of PBL that we

CLARICE WIRKALA is a 2011 graduate of the PhD program in developmental psychology at Teachers College Columbia University; e-mail: crw2115@columbia.edu. Her primary interests are in inquiry and problem-based learning. The research reported here is based on her doctoral dissertation.

DEANNA KUHN is professor of psychology and education at Teachers College Columbia University, 525 West 120th St., New York, NY 10027; e-mail: *Dk100@columbia.edu*. Her current work focuses on design and evaluation of curricula to develop reasoning skills in middle schoolers.

employ in this investigation. These features distinguish PBL from other related instructional methods that are not necessarily problem focused, such as project-based learning (in which the product is a project rather than a problem solution), inquiry learning, and cooperative learning. In the course of addressing such problems, it is expected that students will acquire targeted understanding and knowledge and possibly more general problem-solving skills as well.

Critics have claimed that minimally guided instructional approaches (as they characterize PBL) ignore the structures that make up human cognitive architecture and thus are less effective than instructional approaches that provide greater guidance (Kirschner, Sweller, & Clark, 2006; Mayer, 2004). Particularly in the case of students with little prior knowledge, PBL critics see less guided approaches as less effective because novices may lack the schemas and differentiated knowledge structures needed to incorporate new information into existing knowledge structures (Kirschner et al., 2006). These claims are in direct contrast to those of theorists who maintain that activation of prior knowledge and elaboration are essential to PBL precisely because of its compatibility with human cognitive architecture (Hmelo-Silver, Duncan, & Chinn, 2007; Schmidt, 1993; Schmidt, Loyens, van Gog, & Paas, 2007). The disagreement rests perhaps on whether PBL in fact should be regarded as a minimally guided approach. Far from being "unstructured," its advocates claim, good PBL instruction requires complex, carefully designed instructional protocols, including well-designed scaffolding during each stage of the process (Davies, 2000; Hmelo-Silver et al., 2007). Schwartz and Bransford (1998) and Schwartz and Martin (2004) in fact advocate a method that begins with one or more problems but integrates segments of direct instruction at specific junctures at which students have gained sufficient experience to make use of it.

Most research on PBL has been conducted with adults, most often in medical school settings, comparing the impact of PBL versus traditional, lecture-based curricula on students' knowledge acquisition, clinical performance, and problem-solving skills. Such research has yielded mixed results, and meta-analyses suggest that the alleged superiority of PBL is far from established (Albanese & Mitchell, 1993; Berkson, 1993; Colliver, 2000; Dochy, Segers, Van den Bossche, & Gijbels, 2003; Gijbels, Dochy, Van den Bossche, & Segers, 2005; Hmelo-Silver, 2004; Vernon & Blake, 1993). But even more important is a fact that PBL's critics and proponents both admit: Much of the research is deeply flawed. Studies either occur in artificial environments, far removed from the realities of real educational settings, or at the other extreme, classroom studies lack experimental control and suffer from an array of weaknesses including nonrandom assignment of students to PBL and traditional instruction, variations in time and exposure to treatment across often long interventions, and varying instructors and conditions across treatments (Albanese & Mitchell, 1993; Capon & Kuhn, 2004; Colliver, 2000; Hmelo-Silver, 2004; Vernon & Blake, 1993).

In the present work, we follow the studies of Capon and Kuhn (2004) and Pease and Kuhn (2011) in undertaking to study the effectiveness of PBL in a natural instructional setting yet under tight experimental control. We also follow their approach in acknowledging the varying practices that have been characterized as falling under the heading and undertaking to instantiate PBL in its "best practice" form, namely, as its advocates claim it to be most effective. The multifaceted nature of PBL led Pease and Kuhn to investigate experimentally exactly what the effective components of PBL practice are—an investigation we also pursue in the present work. Specifically, we compare not only PBL and lecture/discussion instructional conditions but also two forms of PBL instruction—team and individual—in order to examine whether the effectiveness of PBL is reduced when its social component is subtracted, and hence whether social collaboration is an essential component of the PBL method.

A major difference between the two studies just cited and the present study is that here we investigate PBL among a K–12 population, specifically middle school students, rather than adult students investigated in those two studies and in most PBL research. PBL along with project-based and collaborative learning have been met with enthusiasm among K–12 educators, despite the time, effort, and cost involved in implementation and despite a scarcity of rigorous evidence of their effectiveness. Rigorous evidence regarding PBL's effectiveness is even scarcer for the pre-college population than it is for adult populations. Given its growing use (Mitchell et al., 2005), and the potential for much more widespread use, in K–12 education, the question of whether its benefits justify its demands is thus one of great practical significance. We see the present study as contributing to an essential research base necessary to answer this question.

Method

Participants

Participants were sixth-grade students at an alternative urban public middle school. The school administration had assigned incoming students to three equivalent classes based on gender, ethnicity, standardized test scores, essay responses on the school's admission exam, and previous academic record. All three classes participated, with *N*s of 30, 29, and 31. The student body at this school is highly diverse, of African American, Hispanic, and Caucasian ethnicities, in approximately equally proportions, with a modest predominance of Hispanic students. Students are also very diverse socioeconomically, coming from low, lower-middle, and upper-middle socioeconomic status (SES) families, with 60% qualifying for free or reduced-price lunch. Students also show a wide range of academic performance, from superior to low average, with all functioning at or above grade level as indicated on standardized tests. (No special education students attend the school.)

To establish baseline understanding of the targeted concepts and the equivalency of the two topics, an additional group of 94 students from one grade below at the same school was administered the cued comprehension assessments (described in the following) but otherwise did not participate in the study.

Design

Topics. The instruction covered two topics. All students were instructed in both topics, via either a PBL or a lecture/discussion (LD) method. The content was developmentally and age appropriate for the population. However, it was entirely new to these students and different from the typical content in the course in which the intervention took place (social studies), thus minimizing previous knowledge as a variable factor across participants. Topic 1 was *groupthink*, the faulty decision making that can occur in groups with low cognitive diversity and other characteristics. Topic 2 was learning and memory, particularly how certain study factors affect memory for learned material.

Length of instruction. The instruction for each topic took place during three 40-minute class sessions, over the course of 1½ weeks, thus lasting a total of 2 hours. The entire intervention, including both topics, thus occupied six class sessions and a total of 4 hours. Instruction on the first topic (groupthink) occurred at the end of the sixth-grade year; instruction on the second (memory) at the beginning of the seventh-grade year. The two topics were separated by summer vacation.

Assessment measures. Only long-term learning was assessed, by means of a cued assessment of comprehension and an uncued assessment of application (to a new context), administered approximately 9 weeks after the instruction for the topic ended. Although short-term gains might have been greater in the LD condition, it is enduring learning that is of interest here.

Conditions. A crossed within-subjects design was employed, manipulating two independent variables: instructional format (PBL vs. LD) and grouping condition (PBL-team vs. PBL-individual). Assignment of classes to conditions is shown in Table 1.

This design enabled us to examine the effect of instructional condition for each topic on both comprehension and application measures by means of both between-subject and within-subject comparisons. Specifically, we asked: (a) Does PBL produce superior results to LD? (b) Does PBL-team

Table 1
Study Design

	Topic 1	Topic 2
Class 1	PBL-individual	PBL-team
Class 2	PBL-team	LD
Class 3	LD	PBL-individual

Note. PBL = problem-based learning; LD = lecture/discussion.

produce superior results to PBL-individual (i.e., is social context an essential component of the PBL method)?

Between-subjects analyses compare performance of PBL-team, PBL-individual, and LD students for each topic. Comparisons are performed separately for each topic. Within-subjects analyses (across topics) compare performance of students across the two methods a student experienced. Analyses of Class 1 compare the effectiveness of PBL-individual and PBL-team; analyses of Class 2 compare the effectiveness of PBL-team and LD; analyses of Class 3 compare the effectiveness of PBL-individual and LD. Thus, in addition to between-subject comparisons of instructional method within topic, each student was able to serve as his or her own control in a comparison of two methods across topics. Analyses establishing the equivalence in the difficulty levels of the two topics made these within-subject comparisons possible.

Topic Content

Topic 1 content encompassed several of the factors that can engender groupthink phenomena: cognitive diversity, conformity, social cohesion, diffusion of responsibility, obedience, and group size (Lahey, 2008; Surowiecki, 2005). Topic 2 content encompassed memory strategies: survey, question, recite, review, reduce interference, spaced learning, and associative links (Higbee, 2001; Lahey, 2008; Lorayne & Lucas, 1974). The specific concepts for each topic are presented in Appendices A and B this article (http://aer.sagepub.com/) in the online version of as supplemental data.

The two topics were chosen to be equivalent in difficulty and equally unfamiliar to students, as well as completely independent of one another, to eliminate the possibility of carry-over or learning effects from one topic to the other. Although the memory concept might seem to be more familiar to these students, in fact, prior to the intervention, students showed almost no knowledge of how these techniques should be used or how they actually improve memory. Assessment results to be presented for the no-intervention group indicated that both topics were equally difficult and equally unfamiliar to students.

PBL Problems

PBL instruction centered around resolving a problem based on a real situation. The problem for each topic was written in the form of a letter, in which the writer asked students to help resolve an important matter.

The letter for the groupthink topic, written by a fictional NASA manager, discussed the *Columbia* space shuttle disaster. Noted was the Mission Management Team's disregard of evidence indicating that damage to the shuttle's external fuel tank during take-off could be detrimental during the shuttle's descent through the atmosphere. The letter also presented some of the groupthink dynamics that took place while the team deliberated on the gravity of the situation. Although a few details were added to highlight certain concepts, the letter was largely based on facts that were uncovered during NASA's investigation (Surowiecki, 2005). The letter writer asked specifically for help to solve this mystery:

> Why do you believe the team members didn't collect more information and try to find out everything they could to make the right decision about the *Columbia*? What characteristics about this particular group made it not work well and what could have helped them function better as a group? You can imagine how happy I will be once I resolve this management problem. Indeed, the formation of effective specialized groups is very important for the safety of NASA's astronauts.

The letter for the memory topic was written by a fictional doctor wanting to resolve the crisis of deaths caused by medical error, taken from real cases, by improving medical students' study habits. The letter writer asked for help in improving medical students' study habits, in particular memorization strategies:

> What are they doing wrong in their studying? What advice can you give them on effective studying and memorization skills and why do you believe these will work? What practical strategies should our students and current surgeons use to memorize important procedures, in the right order?

Procedure

To maximize equivalency across groups, the first author taught all LD sessions and was one of three "coaches" in PBL sessions. Table 2 presents an overview of the instructional procedure in the PBL and LD conditions, highlighting their parallels and differences. Details of the procedure in each condition follow.

PBL-individual format. PBL classes were led by three adults, who circulated the classroom and served as "coaches," answering questions and encouraging students to stay on task. Students in PBL-individual classes

Table 2

Overview of Instructional Sequence in Problem-Based Learning (PBL) and Lecture/Discussion (LD) Conditions

Problem-Based (Team and Individual)	Lecture/Discussion
Session 1	
Problem-formulation and initial analysis	
Introduction to topic (5 minutes)[a]	Introduction to topic(5 minutes)[a]
Identify problem and related facts (15 minutes)	Lecture/discussion on first half of concepts (30 minutes)
Address PBL problem (15 minutes)	Introduction to all concepts (5 minutes)[a]
Introduction to all concepts (5 minutes)[a]	
Session 2	
Problem analysis	
Condensed lecture/discussion on all concepts with concepts handout available (20 minutes)	Lecture/discussion on second half of concepts and related concepts, with concepts handout available (40 minutes)
Utilize concepts to solve problem (without concepts handout) (20 minutes)	
Session 3	
Problem resolution	
Final problem resolution, with concepts handout available (35 minutes)	Lecture/discussion on all concepts, with concepts handout available and new examples(35 minutes)
Concluding discussion (5 minutes)[a]	Concluding discussion (5 minutes)[a]

Note. Concepts handout was similar to that shown in Appendices A and B found as supplemental data to this article online.

[a]Segment is the same for both PBL and lecture classes.

were instructed to work alone and engage in no talk with classmates. Students were presented the problem and spent their time working quietly on it, while the coaches answered occasional questions if students raised their hands. The only exception to the quiet atmosphere of individual work was during the condensed lecture, when a few students had the chance to ask questions about the material in front of the class. To facilitate their work, students were provided the same scaffolding handout (see the following) provided to the PBL teams.

PBL-team format. PBL-team classes were also led by three coaches. Students in the PBL-team condition were randomly assigned to a team, each with a "team leader" who was appointed by the instructor (based on reputation for good behavior and focus). Students were presented the problem and told they needed to discuss it and come up with a problem resolution *as a team* and that they should take turns writing down responses on a scaffolding handout (see the following). While some teams were very capable collaborators, with all three members actively referring to the letter and engaging in sometimes heated debates about the problem, other teams relied on one or two members to do most of the work, with the rest relaxing and saying little. Teams showed varying levels of collaboration, with most teams collaborating relatively well, especially when encouraged to do so. Thus, one of the main jobs of the coaches in the PBL-team classes, in addition to answering questions, was to remind students that everyone needed to contribute in order to come up with the best problem resolution and that they could not simply rely on the efforts of any one student. During the final, problem resolution phase (Session 3, Table 2), PBL-team students worked on the problem resolution individually for 15 minutes (the only time segment in which they worked by themselves in this condition) and then in their teams for the remaining 20 minutes. This ensured that all students struggled with the problem on their own.

PBL instructional sequence. Except for the grouping condition, PBL-team and PBL-individual classes were conducted in an identical manner, engaging in exactly the same tasks. During Session 1, the problem formulation and initial analysis stage, students were introduced to the topic as an "advanced social sciences unit" taking place in their social studies class for the next few class periods. Students were told, "There will be a pop quiz on this material later, so it's important that you pay close attention." Students were then presented with the letter, and using a handout as scaffolding, they were asked to identify: the main problem, questions they need to look into to answer the problem, facts presented in the letter that could possibly help them solve it, and what else they might need to find out about in order to solve it. Students then generated hypotheses regarding how to address the problem and constructed a preliminary answer to the letter. In contrast to standard classroom activities, in which students are first presented with new concepts and are subsequently asked to employ what they have learned, often by answering comprehension questions, PBL students thus engaged the problem "cold," without having been introduced to relevant concepts. Thus, at this early stage, students needed to rely on their preexisting knowledge to address the problem. In the last 5 minutes of Session 1, students were briefly introduced to all relevant concepts and terms.

In the majority of cases, students came up with elaborated but overly simplistic problem resolutions during Session 1, such as: "Write the checklist on your hand," "Get better medical students," and "The NASA people are just lazy so you should fire them." Thus, it was announced at the beginning of Session 2 that coaches had read their resolutions from Day 1, and that although there were some good efforts, most people had not addressed the heart of the problem and were looking for "easy answers."

During Session 2, the problem-analysis stage, students learned about the relevant concepts via a condensed, 20-minute lecture, while looking at a handout with the concepts briefly defined (as in Appendices A and B) as a visual aid. Although this lecture covered the concepts and theories pertinent to the topic, neither the letter nor the problem was explicitly addressed. Students were simply told, "We're going to talk about some things that might help you to solve this problem." The groupthink lecture focused on the psychologists, experiments, and research of group dynamics (e.g., Asch, Milgram, Zimbardo). The memory lecture explored study skills and memory strategies by focusing on the results of relevant research studies and experiments. The condensed lecture in the PBL condition differed from that in the LD condition in that it was much more limited in length and scope. PBL students were provided the same essential information as LD students, but in telescoped fashion.

Because only 20 minutes were allotted to this activity, the lecture succinctly covered only the essentials, briefly describing significant experiments and their conclusions, providing brief definitions and explanations of concepts, as well as one or two examples of possible applications. Although many students wanted to make comments and ask questions during the lecture, there was time for only a couple of students to ask questions before the class was asked to move on to their problem resolution.

For the remainder of the class, students worked toward a solution to the problem. The concepts handout that students looked at during the lecture was collected, so students could not use these as a crutch while solving the problem at this stage. Whether or not they drew on the concepts introduced, students needed to make the intellectual effort of constructing a problem resolution. Because the objective was to move students away from preexisting theories and toward the new concepts, they did not have access to the problem solutions that they generated during Session 1.

Problem resolutions were more advanced after Session 2. For example, some students' resolutions revolved around the fact that NASA "didn't have the right information" or that the group "didn't function well because they had never worked outside of NASA." Coaches responded to and scaffolded such problem resolutions by asking students, for example, "Why didn't they get the information?" or "Why would not working outside of NASA mean they didn't function well?" (This response would direct them to the concept of cognitive diversity, which they would eventually learn about or would already have heard about in the condensed lecture.)

Students spent most of Session 3 working through the problem for a final time and generating their resolution in the form of a letter. They did not have access to their problem solutions from Session 2 but were given the one-page concepts handout. Students were told that although their problem resolutions from Session 2 were an improvement on those from Session 1, they were still not adequate, and they needed to utilize the concepts in order to rethink, refine, and extend their problem solutions. Finally, the last 5 minutes of Session 3 were spent on discussion of students' problem resolutions.

Student questions and role of coaches in PBL condition. In both PBL-team and PBL-individual classes, coaches were available to answer questions but did not offer any "answers" to students; they only addressed questions and confusions when they were specifically asked. About one-third of students actively sought out help in each class, and the same types of questions were asked in PBL-team and PBL-individual classes. Students primarily asked procedural questions, such as, "Do I have to fill out this whole sheet?" "Can I take it home?" and "Will this be graded?" Many questions were for clarification of the problem, such as, "What does fuselage mean?" "Is the MMT the same thing as the astronauts?" and "Did these medical mistakes really happen?" Content questions were also asked (but to a lesser extent than procedural and clarification questions), such as, "Why didn't they ask questions during the team meeting?" (To answer this question, coaches responded, "We don't know; that's for you to think about.") "Did the NASA people think the shuttle was going to be okay?" ("Read the letter to find out.") "How could they amputate the wrong leg?" ("I don't know—but what's the main problem that you need to solve in this letter?"). Sometimes students asked content questions in which they were clearly trying to circumvent the main problem or find an "easy answer," such as, "Are these medical students dumb?" ("These schools are very selective; read the letter.") and "Are they talking about really bad hospitals?" ("No, medical error happens in all hospitals, including the best ones.").

Most of the coaches' efforts revolved around encouraging students to re-read the letter carefully and identify relevant facts. While many of the more capable students underlined key parts of the letter as they read, others appeared overwhelmed by the information in the letter. Another of the coaches' jobs was to help students understand what the problem truly was. For example, some students at least initially thought that the groupthink problem was about fixing the shuttle so that it would not explode rather than figuring out what went wrong in the group dynamics. Likewise, many students initially thought that the memory problem they needed to deal with was the medical errors themselves rather than what was going wrong with the students' studying and how their study skills could be improved. Coaches encouraged these students to underline the main question that the writer was asking and then go back and underline any facts that they felt might be relevant in solving it. Thus, students were encouraged to read and re-read the letter carefully, keeping certain questions in mind. Students were never told what the "problem" or relevant facts were, what to write on their worksheets, or how to solve the problem.

Lecture/discussion instructional sequence. During Session 1, as in the PBL class, students in the LD condition were introduced to the topic as an "advanced social sciences unit" taking place during the next few class periods. They were also told: "There will be a pop quiz on this material later, so it's important that you pay close attention." Subsequently, most of the class was spent on lecture and discussion on the first half of the concepts covered in that topic. The lecture covered the same concepts as in the PBL condition (Appendices A and B) but made no mention of the problem from the PBL condition. Care was taken in planning and executing the LD instruction to cover all concepts and give them equal weight in instruction. The critical difference between PBL and LD conditions, note, is that PBL students had to address and solve a problem, while LD students participated in lecture and discussion only and were not asked to solve a problem (although LD material used to illustrate points did give students some exposure to such problems).

Whereas PBL students had only the essential concepts defined and briefly explicated in the condensed lecture, LD students heard and discussed each of the concepts in detail, thoroughly exploring relevant theories and research. Moreover, students in the LD condition had the benefit of being exposed to multiple examples and various applications of each of the concepts, considered to be an effective way to promote learning (Van Merrienboer & Sweller, 2005). Each new concept was explicitly contrasted to the other concepts, so that students' attention was directed to their differences and similarities, in this way continually reviewing and refining their understanding of previously learned concepts. Further differentiating the LD and PBL conditions, students in the LD group had ample time to make comments, ask questions, and clarify their confusions. Discussion centered around students' questions and comments. At the end of Session 1, as in the PBL condition, students were briefly introduced to all of the concepts for the topic, serving as an introduction to the concepts that had yet to be covered.

Session 2 was spent entirely on lecture and discussion of the other half of the concepts that were not covered during the first session as well as coverage of related concepts that deepened understanding of the main topic. (For example, for the memory topic, students heard about the difference between short-term and long-term memory, deepening their understanding of "reviewing" and its benefits.) As a visual aid during Sessions 2 and 3, students looked at the same concepts handout used by PBL students, briefly defining the concepts.

During the last session, students had the opportunity to refine their understanding of all of the concepts with new, extended examples. Finally, as in the PBL conditions, the last 5 minutes of the class were spent on a reflective discussion of the main topic.

Student questions and role of coaches in LD condition. Because in the LD classes all questions were addressed to the entire class and there was ample time for all students to voice their questions and comments, these were staffed by only two adults. One led the lecture/discussion while the other circulated the classroom. LD students did not take notes during the lecture (nor did PBL students take notes during the condensed lecture). (Note taking was not included as a component of the instruction for several reasons. Students of this age are not effective note takers. Moreover, doing so could have hurt LD students by taking their attention away from the teacher and the class discussion. Also, we would not have wanted students to make use of such notes outside of class—the customary purpose of note taking—as this would have confounded the experimental comparison, given that PBL students were not spending any extra outside of class time.)

Students in the LD classes asked several clarification questions, such as (referring to the Milgram experiments on obedience) "Did those people really think that they were shocking another person?" or (referring to spaced vs. massed learning) "Why do they always use rats in experiments?" Students also asked deeper content questions that clarified confusions they had about the material, such as, for groupthink, "What is the difference between homogeneous and heterogeneous groups?" or "How can people conform without knowing they are conforming?" Content questions for memory were, for example, "Why is linking a good technique if it's more work?" or "Why are images more memorable than words?" These questions were sometimes put to other students to try to answer, in order to engage more of the class in the dialogic manner typical of classroom teachers, but they were always fully answered by the lecturer. Most comments made by students related the material to their personal experience. For example (regarding the Asch conformity experiments), one student said, "People at school always try to be like everyone else by dressing the same." During the memory topic, students raised their hands to share their own learning and study tactics.

Students were encouraged to engage in and discuss the lecture material so their attention would not wander through passive listening. Most students participated at least once per class, and approximately half of the students made comments and asked questions multiple times during each class. During the memory lecture, students were asked to practice some of the techniques they were taught, such as linking various words together in a story, and to share their links with the class. To illustrate the concepts, they were asked other questions, such as, "How many windows are in your home?" and "How did you come up with this number?" (This led to the discussion on the power of visual imagery in memory.) Students were engaged in the groupthink topic by being asked specific questions that led up to the concepts, such as, "Let's say that you want to solve the problem of how to improve students' grades in schools. What kind of people would you ask to help you solve this problem?" (Students' responses led to a discussion of whether it was better to have all principals in their task force, in a homogenous group, or a group of principals, teachers, and students, in a cognitively diverse group.) To ensure that the lecture was as engaging as possible, there was no predetermined or rigid time frame for the different activities. Rather, class was conducted as a dynamic back-and-forth between lecture, examples, questions and answers, and discussion.

Assessments

The assessments took place approximately 9 weeks after the end of the instruction, during a single class period. All students worked individually and were not allowed to use any notes or other aids, given the purpose of the assessments was to measure recall, in addition to comprehension and application of concepts. This lapse of time ensured that deep learning was being assessed because students needed to understand and integrate the new knowledge in order for it to be retained in long-term memory. There were two types of assessments. One was a *comprehension* assessment in which the concepts were directly cued (named in a provided list) and the student was asked to explain each of them. The other was an indirectly cued assessment of *application*, in which students were presented a new situation to which the concepts could be applied but the concepts were not identified. To ensure that there was no priming effect, the application assessment was given before the comprehension assessment. Students were given 20 minutes to work on each assessment.

Application assessment. The application assessment measured students' integration and application of the concepts to a new context. The main topic and concepts were not referred to in the essay question, and students could potentially respond without mentioning the concept. Therefore, to recognize their applicability in the new context, students needed to have understood, retained, and integrated these concepts into their long-term knowledge structures. Thus, this assessment tested deep understanding by examining whether students *spontaneously* applied the concepts learned to a novel situation, without being explicitly prompted to do so.

The following was the groupthink application assessment:

> You are President Obama's head diplomat to Iran. Iran has a nuclear energy program and it's possible that they already have a nuclear bomb. Iran's president has expressed hostility toward the US, so diplomatic efforts must be handled with maximum competence and intelligence. Obama has made you director of a committee to plan negotiations with Iran. How would you select and run your committee, to make sure it is successful? Give as detailed an answer as you can.

The application assessment questions were significantly different from the problem that the PBL students addressed. In the groupthink problem, students focused on *what went wrong* with the NASA group, but in this assessment, they needed to think about how a group could be run successfully. Although students could use the basic groupthink concepts to respond to the question, they needed to redirect their focus from the causes of groupthink to how it could be prevented. This had been only briefly considered in the last instructional session, to an equal extent in both PBL and LD groups (e.g., how to combat conformity, obedience, etc.).

The application assessment for the memory topic similarly was significantly different from the problem that PBL students had addressed. Both the PBL and LD instruction focused on learning and memory skills and strategies. The memory assessment problem focused on a unique scenario that had the potential to lead them astray. Therefore, as in the groupthink assessment, students needed to have understood, retained, and integrated the instructional concepts into their long-term knowledge structures in order to recognize their applicability to this novel case.

This was the memory application assessment:

> You are a *New York Times* journalist, and you will be traveling around the world for the next 3 months—undercover—to learn about terrorism. You are interested in answering 3 main questions related to terrorism: What do people from other cultures admire about the United States? What do people from other cultures dislike about the United States? How do terrorist groups recruit new members? After the trip, you will write a front-page article, reporting the most important things you learned. To keep your identity a secret, you will not write down or tape record anything while you interview people, but will take some notes at the end of each day when you're back in the hotel. When the 3 months are over and you are going home you will have to destroy ALL of your notes—some of which contain important, secret material—in case your luggage is searched at the airport. The penalties are severe, so you don't want to try anything tricky to get your written notes back home. You're going to have to rely on what you can keep in your head. How can you make sure that the article you write when you get home is as accurate as possible, reflecting everything you learned? Give as detailed an answer as you can.

Comprehension assessment. The comprehension assessment measured students' comprehension of the concepts. The instructions for the groupthink and memory topics were as follows:

Please define and fully explain the following groupthink concepts:

1. Groupthink
2. Cognitive diversity
3. Conformity
4. Social cohesion
5. Diffusion of responsibility
6. Obedience
7. Group size

Please define and fully explain the following memory concepts:

1. Survey
2. Question
3. Recite

4. Review
5. Reduce interference
6. Space learning
7. Associative links

Results

Coding

Comprehension assessment. A two-tiered coding system was employed for the cued comprehension assessment. The first tier addressed whether a concept was defined at all. For each topic, students were given 1 point for each of the seven concepts that they correctly defined in at least a basic way, and these points were summed to create a total score for definition. Total possible score was thus 7 for each topic.

In a second tier of coding, for each concept a student defined, a score was assigned for the level of explanation achieved, based on this ordinal scale:

0 = No relevant response;
1 = Basic definition: provides only vague or very basic definition;
2 = Elaborated definition: provides basic definition and elaborates on definition;
3 = Basic explanation: provides basic definition, elaborates on definition, and provides basic explanation;
4 = Elaborated explanation: provides basic definition, elaborates on definition and provides basic explanation; also elaborates on explanation or relates the concept to the main topic or related concepts.

The coding levels are cumulative, so each level presupposes a positive score on all lower levels. For each topic, the *highest* level that the student attained for any concept was noted. The modal level of explanation across all concepts the student defined for that topic was also identified. (Where there was no mode, the median was noted.) Examples of student responses at each level for the groupthink topic appear in Appendix C and for the memory topic in Appendix D (see online supplemental data).

Application assessment. A two-tiered coding system was similarly employed for the application assessment. First, for each topic students were given 1 point for each of the seven concepts they invoked in their essay. Total possible score was thus 7 for each topic.

In a second tier of coding, for each concept a student identified, a score was assigned for the level of explanation achieved, based on this ordinal scale:

0 = No reference: does not apply the concept, either in content or by name;
1 = Mention: applies the concept, either in content or by name, but does not define;

2 = Definition: applies the concept, either in content or by name, and provides definition;
3 = Explanation: Applies the concept, either in content or by name, and provides definition and explanation;
4 = Elaborated explanation: Applies the concept, either in content or by name, provides definition and explanation, and also elaborates on the application or explanation or establishes meaningful relationships between concepts.

For each topic, the *highest* level that the student attained for any concept was noted, assessing students' ability not only to integrate the defined concept into a meaningful knowledge structure but to apply this knowledge to a new context in which it had the potential to be useful. Because the questions did not directly draw students' attention to the concepts and students could ignore the new concepts in responding to the application assessments (as many did), modal levels of explanation were not identified, since they were often zero, even in cases of high-quality responses in which a student ignored several concepts while successfully applying several others (e.g., the student could apply the concepts at an explanation level of 0, 2, 4, 3, 0, 4, 3, 0 and their modal level would still be zero). Examples of student responses at each level for the groupthink topic appear in Appendix E and for the memory topic in Appendix F (see online supplemental data).

Coding Reliability

A primary coder (the first author) coded all responses, and a proportion (20%) of responses across topics and conditions was coded by a second coder who was a trained doctoral student but not otherwise involved in the study. Both coders were blind to the students' identity and condition, as well as the other coder's scores. Percentage agreement between the two coders was as follows: groupthink comprehension, 86% (Cohen's kappa = .82); groupthink application, 86.8% (Cohen's kappa = .79); memory comprehension, 91% (Cohen's kappa = .87); memory application, 89% (Cohen's kappa = .70).

Statistical Analysis

In order to address the research questions of specific interest in this study, analyses for most performance variables consisted of two planned comparisons—one between the LD group and the PBL groups and one between the two PBL groups. A *t* test was used to assess the difference in mean number of concepts defined/applied. The chi-square statistic was used where ordinal scales were involved to assess depth of explanation achieved. The Wilcoxon signed-rank test, a nonparametric test used for repeated measures analyses, was used to assess individual patterns over the two topics. (The Wilcoxon uses the *Z* statistic and evaluates differences between repeated scores based on the magnitude of the difference between pairs of observations. Statistical significance was set at the .05 alpha level.) Students who were absent for any of the three sessions for a topic were not included in analyses. Across topics, 24 to 31 students were included in the between-subjects analyses, and 19 to 22 were included in the within-subjects analyses.

No-Instruction Baseline Performance

Results of comprehension assessments for the two topics among the no-instruction group indicated that students had negligible prior knowledge of the concepts. These results also established that the two topics were equivalent in difficulty. Mean number of concepts the no-instruction group defined for the groupthink topic was .56 (*SD* = .80; range 0–3). Mean number of concepts the no-instruction group defined for the memory topic was .52 (*SD* = .90; range 0–5). The difference across topics was not significant, *t*(186) = .34, *p* =.732. The modal number of concepts defined for each concept was zero, and most students who were able to define any concept only defined one. The few students who were able to define any one concept provided only a very basic definition, with no student reaching the explanation level. A comparison of the highest level of explanation achieved for any concept across topics further supported the equivalency of the two topics, χ^2(2, *N* = 188) = 2.13, *p* = .346.

Performance Following Instruction for Groupthink Topic

Comprehension. Initial comprehension prior to instruction was not assessed for the main experimental groups, so as not to prime their use of the concepts or otherwise interfere with the effects of instruction. Following instruction, for the groupthink comprehension assessment, mean number of concepts defined by the PBL-team group was 5.19 (*SD* = 1.33; range 2–7). Mean number of concepts defined by the PBL-individual group was 4.84 (*SD* = 1.18; range 2–7). The difference between the two PBL groups in mean number of concepts defined was not significant, *t*(52) = .99, *p* = .329. Similarly, the difference in modal explanation levels reached by PBL-team versus PBL-individual groups was only of marginal significance, χ^2(2, *N* = 52) = 7.89, *p* = .050 (with PBL-team tending toward higher performance). Nor was there a significant difference in the highest level of explanation reached, χ^2(2, *N* = 52) = 2.21, *p* = .331 (although in this case PBL-individual tended toward higher performance). (See Tables 3 and 4.)

PBL groups, however, defined a higher mean number of concepts than did the LD group. Mean number of concepts defined by the combined PBL groups was 5.02 (*SD* = 1.26; range 2–7). Mean number of concepts defined by the LD group was 3.33 (*SD* = 1.37; range 1–6). This difference between the combined PBL groups and the LD group was significant, *t*(74) = 5.27, *p* = .001. The combined PBL group also showed higher levels of explanation

Table 3
Percentage of Students in Each Category Based on Highest Explanation Level Given in Groupthink Comprehension Assessment

	Elaborated Definition or Lower	Basic Explanation	Elaborated Explanation
PBL-team	29.6	33.3	37.0
PBL-individual	16.0	52.0	32.0
LD	66.6	25.0	8.3

Note. For the statistical comparison of PBL-team versus PBL-individual, the five original coding levels were collapsed into three levels to obtain higher cell values for the chi-square test. The first three coding levels (no relevant response, basic definition, and elaborated definition) were combined into one lower level, while basic explanation and elaborated explanation comprised the two highest levels. PBL = problem-based learning; LD = lecture/discussion.

Table 4
Percentage of Students in Each Category Based on Modal Explanation Level Given in Groupthink Comprehension Assessment

	No Relevant Response	Basic Definition	Elaborated Definition	Basic or Elaborated Explanation
PBL-team	18.5	33.3	25.9	22.2
PBL-individual	44.0	8.0	36.0	12.0
LD	70.8	16.7	12.5	0

Note. Where there was no mode, the median was noted. For statistical comparison of PBL-team versus PBL-individual, as well as PBL-team and PBL-individual versus LD, the five modal levels that students reached were collapsed into four levels to obtain higher cells values for the chi-square test. The two highest levels, 3 and 4, were combined into one level, and these were compared to the lower three modal levels of 0, 1, and 2. PBL = problem-based learning; LD = lecture/discussion.

than the LD group, $\chi^2(3, N = 76) = 15, p = .002$. PBL students were overrepresented at the highest levels and underrepresented at the lower levels; the majority of PBL students reached the explanation levels, while the majority of LD students reached only the definition levels (Table 3). Finally, the combined PBL group also showed higher modal explanation levels than the LD group, $\chi^2(3, N = 76) = 12.58, p = .006$. Nearly half of PBL students reached modal levels of 2 or above, while the vast majority of LD students only reached modal levels of 0 and 1 (Table 4).

Application. For the groupthink application assessment, mean number of concepts applied by the PBL-team group was 2.59 (*SD* = 1.82; range 0–6). Mean number of concepts applied by the PBL-individual group was 2.88 (*SD* = 1.97; range 0–6). The difference between PBL groups in mean number of concepts applied was not significant, $t(51) = -.56, p = .577$. Similarly, there

Table 5

Percentage of Students in Each Category Based on Highest Explanation Level Given in Groupthink Application Assessment

	No Reference or Mention Only	Definition	Explanation	Elaborated Explanation
PBL-team	18.5	25.9	25.9	29.6
PBL-individual	15.4	30.8	38.5	15.4
LD	58.3	25.0	8.3	8.3

Note. For statistical comparison of PBL-team and PBL-individual versus LD, the five original coding levels were collapsed into four levels to obtain higher cells values for the chi-square test. The first two coding levels (no reference and mention) were combined into one level lower, while definition, explanation, and elaborated explanation comprised the last three levels. PBL = problem-based learning; LD = lecture/discussion.

was no significant difference in the highest level of explanation reached by PBL-team versus PBL-individual groups, $\chi^2(3, N = 53) = 2.02, p = .568$.

The combined PBL group, however, applied a higher mean number of concepts than did the LD group. Mean number of concepts applied by the combined PBL group was 2.74 (*SD* = 1.88; range 0–6). Mean number of concepts applied by the LD group was 1.17 (*SD* = 1.52; range 0–5). The difference between the combined PBL group and the LD group was significant, $t(74) = 5.27, p = .001$. The combined PBL group also showed higher levels of explanation than the LD group, $\chi^2(3, N = 77) = 15.16, p = .002$. As in the comprehension assessment, PBL students in the application assessment were overrepresented at the highest levels and underrepresented at the lower levels; the majority of students in the LD group performed at the no reference, mention, and definition levels, while the majority of PBL students reached the explanation levels (Table 5).

Performance Following Instruction for Memory Topic

Comprehension. Following instruction, for the memory comprehension assessment, mean number of concepts defined by the PBL-team group was 4.70 (*SD* = 1.60; range 2–7). Mean number of concepts defined by the PBL-individual group was 4.00 (*SD* = 1.71; range 0–7). The difference between the two groups in mean number of concepts defined was not significant, $t(59) = –1.65, p = .105$. There was also no significant difference in the highest explanation levels reached by PBL-team versus PBL-individual groups, $\chi^2(3, N = 61) = 1.24, p = .744$ (Table 6), nor in the modal level of explanation reached, $\chi^2(3, N = 61) = 1.15, p = .765$ (Table 7).

The combined PBL group, however, defined a higher mean number of concepts than did the LD group. Mean number of concepts defined by the combined PBL group was 4.34 (*SD* = 1.68; range 0–7). Mean number of

Table 6
Percentage of Students in Each Category Based on Highest Explanation Level Given in Memory Comprehension Assessment

	No Response or Basic Definition Only	Elaborated Definition	Basic Explanation	Elaborated Explanation
PBL-team	13.3	20.0	40.0	26.7
PBL-individual	9.7	32.3	35.5	22.6
LD	14.3	60.7	10.7	14.3

Note. For comparison of PBL-team versus PBL-individual as well as PBL-team and PBL-individual versus LD, the five original coding levels were collapsed into four levels to obtain higher cells values for the chi-square test. The first two coding levels (no relevant response and basic definition) were combined into one lower level, while elaborated definition, basic explanation, and elaborated explanation comprised the three highest levels. PBL = problem-based learning; LD = lecture/discussion.

Table 7
Percentage of Students in Each Category Based on Modal Explanation Level Given in Memory Comprehension Assessment

	No Response or Basic Definition Only	Elaborated Definition	Basic Explanation	Elaborated Explanation
PBL-team	50.0	30.0	20.0	0
PBL-individual	61.3	19.4	19.4	0
LD	85.7	14.3	0	0

Note. PBL = problem-based learning; LD = lecture/discussion.

concepts defined by the LD group was 2.75 (*SD* = 1.40; range 1–6). The difference between the combined PBL group and LD group was significant, $t(87) = 4.36$, $p < .001$. The combined PBL group also showed higher levels of explanation than the LD group (Table 6), $\chi^2(3, N = 89) = 12.02$, $p = .007$. The majority of PBL students reached the explanation levels, while the majority of LD students reached only definition levels. The combined PBL group also showed higher modal explanation levels than the LD group (Table 7), $\chi^2(3, N = 89) = 15.67$, $p = .001$.

Application. For the memory application assessment, mean number of concepts applied by the PBL-team group was 2.10 (*SD* = 1.75; range 0–6). Mean number of concepts applied by the PBL-individual group was 2.42 (*SD* = 1.75; range 0–7). The difference between groups in mean number of concepts applied was not significant, $t(59) = -.71$, $p = .478$. Similarly, there was no significant difference in the highest level of explanation reached by PBL-team versus PBL-individual groups, $\chi^2(4, N = 61) = 4.60$, $p = .331$.

Table 8

Percentage of Students in Each Category Based on Highest Application Level Given in Memory Application Assessment

	No Reference or Mention Only	Definition	Explanation	Elaborated Explanation
PBL-team	46.6	3.3	33.3	16.7
PBL-individual	32.3	19.4	29.0	19.4
LD	65.5	20.7	6.9	6.9

Note. PBL = problem-based learning; LD = lecture/discussion.

The combined PBL group, however, applied a higher mean number of concepts than did the LD group. Mean number of concepts applied by the combined PBL group was 2.26 (*SD* = 1.74; range 0–7). Mean number of concepts applied by the LD group was 1.24 (*SD* = 1.15; range 0–5). The difference between the combined PBL group and the LD group was significant, $t(78) = -3.30$, $p = .001$. (Levene's test indicated unequal variances, $F = 7.28$, $p = .008$, so degrees of freedom were adjusted from 88 to 78.) The combined PBL group also showed higher levels of explanation than the LD group (Table 8), $\chi^2(4, N = 90) = 11.31$, $p = .023$. Nearly half of PBL students reached the explanation levels, while the majority of LD students reached only definitional levels.

Comparison Across Topics

Especially because participants encountered the two topics at two distinct times separated by several months, in the analyses presented thus far we elected to analyze the two topics separately, in effect treating one as a replication of the other to establish that results were not specific to one topic. However, it was also of interest in a secondary set of analyses to examine each group's performance across topics for the two conditions they encountered. To do so, it is essential to establish that the topics are of equivalent difficulty. As reported earlier, we did this for an independent sample. However, it is also desirable to do so for the main sample in the assessments that followed instruction. Comparisons across the two topics for all performance assessments were consistent with the results reported for the noninstruction group: The two topics were of equivalent difficulty. All but one comparison were nonsignificant at the .05 level. We present those results here.

Comprehension. For the groupthink comprehension assessment, mean number of concepts defined was 4.49 (*SD* = 1.51; range 1–7). For the memory comprehension assessment, mean number of concepts defined was 3.84 (*SD* = 1.76; range 0–7). This was the only comparison that was significant, $F(1, 163) = 6.26$, $p = .03$. A comparison of the modal explanation level for groupthink comprehension versus memory was nonsignificant, *F*(1, 163) = .68, *p* = .41, as was the highest level of explanation reached for the two topics, *F*(1, 163) = 2.26, *p* = .135.

Application. For the groupthink application assessment, mean percentage of concepts applied was 28.08% (*SD* = 23.92; range 0%–75%). (Percentages are used for this comparison because, cognitive diversity was split into two concepts, making 8 vs. 7.[1]) For the memory application assessment, mean percentage of concepts applied was 27.62% (*SD* = 23.44; range 0%–100%). This difference was nonsignificant, *F*(1, 165) = .02, *p* = .899. The highest level of explanation for groupthink versus memory application was also nonsignificant, *F*(1, 165) =1.19, *p* = .277.

Within-Group Analyses of Individual Patterns

The overall equivalence of difficulty level across topics permitted an additional set of analyses to be conducted within groups across the two instructional methods that the group experienced.[2]

PBL-individual versus PBL-team comprehension. In Class 1, students studied the groupthink topic via PBL-individual and the memory topic via PBL-team. In the comprehension assessment, mean number of concepts defined when this group's learning occurred via PBL-individual was 4.90 (*SD* = 1.18; range 2–7), and mean number of concepts defined when it occurred via PBL-team was 4.62 (*SD* = 1.43; range 2–7). This difference was nonsignificant, *Z* = –.96, *p* = .339. Nine students defined more concepts when learning through PBL-individual, 7 defined more concepts when learning through PBL-team, and 5 students defined the same number of concepts in both. There was also no significant difference in modal explanation levels, *Z* = –.40, *p* = .689, nor highest level of explanation (*M*s = 3.10 and 2.81) when learning through PBL-individual versus team, *Z* = –1.51, *p* = .130. Seven students achieved a higher level of explanation when learning via PBL-individual, 3 students achieved a higher level of explanation when learning via PBL-team, and 11 students achieved the same level of explanation in both.

PBL-individual versus PBL-team application. As noted earlier, percentages were used since there were eight concepts for groupthink and seven concepts for memory ("cognitive diversity" was split into two concepts). Mean percentage of concepts applied when learning occurred via the PBL-individual method was 39.77% (*SD* = 23.98; range 0%–75%), and mean percentage of concepts applied when learning occurred via the PBL-team method was 30.52% (*SD* = 23.03; range 0%–71%). This difference was nonsignificant, *Z* = –1.54, *p* = .123. Fifteen students applied more concepts when learning via PBL-individual, and 7 students applied more concepts when learning via PBL-team. There was also no significant difference in highest level of explanation achieved (*M*s = 2.59 and 2.09), *Z*

= –1.78, p = .075. Eight students achieved a higher level of explanation when learning via PBL-individual, 5 students achieved a higher level of explanation when learning via PBL-team, and 9 students achieved the same level of explanation in both.

Analyses of individual patterns thus confirm the results of the between-subjects comparisons: Neither comprehension nor application differs significantly across PBL-individual and PBL-team instructional conditions.

PBL-team versus LD comprehension. Students in Class 2 learned groupthink concepts via the PBL-team method and memory concepts via the LD method. Mean number of concepts defined when learning occurred via PBL-team was 4.95 (*SD* = 1.32; range 2–7), and mean number of concepts defined when learning occurred via LD was 2.67 (*SD* = 1.49; range 1–6). This difference was significant, Z = –3.73, p < .001. Eighteen students defined more concepts when learning via PBL-team, only 1 student defined more concepts when learning via LD, and 2 students defined the same number of concepts in each. The majority of students also reached higher modal levels of explanation when learning took place via PBL-team compared to LD, Z = –3.48, p < .001. The majority of students had modal levels of 2 or above when learning via PBL and 0 when learning via LD. Most students also reached higher levels of explanation when learning via PBL (*M*s = 2.90 and 2.24), Z = –2.57, p = .01. Twelve students achieved a higher level of explanation when learning via PBL-team, 3 students achieved a higher level of explanation when learning through LD, and 6 students achieved the same level of explanation in each.

PBL-team versus LD application. For the application assessment, mean percentage of concepts applied when learning via PBL-team was 35.80% (*SD* = 22.26; range 0%–75%), and mean percentage of concepts applied when learning via LD was 20.78% (*SD* = 16.92; range 0%–71%). This difference was significant, Z = –2.14, p = .032. Sixteen students applied more concepts when learning via PBL-team, 5 students applied more concepts when learning via LD, and one student applied the same number of concepts in each. The majority of students also reached a higher level of explanation when learning took place via PBL-team compared to LD (*M*s = 2.73 and 1.45), Z = –2.963, p = .003. Fourteen students achieved a higher level of explanation when learning via PBL-team, 3 students achieved a higher level of explanation when learning through LD, and 5 students achieved the same level of explanation in each.

Analyses of individual patterns in Class 2 thus also confirm the results of the between-subjects comparisons. The majority of students recalled, comprehended, and applied concepts better when learning took place via PBL compared to LD.

PBL-individual versus LD comprehension. Students in Class 3 learned groupthink concepts via LD and memory concepts via PBL-individual. For the comprehension assessment, mean number of concepts defined when learning via PBL-individual was 4.32 (*SD* = 1.73; range 0–7), and mean number of concepts defined when learning via LD was 3.42 (*SD* = 1.50; range 1–6). This difference was significant, *Z* = –2.11, *p* = .035. Ten students defined more concepts when learning via PBL-individual, only 2 students defined more concepts when learning through LD, and 7 students defined the same number of concepts in each. The majority of students reached higher modal levels of explanation when learning took place through PBL-individual compared to LD, *Z* = –2.25, *p* = .024. The majority of students had modal levels of 1 or above when learning through PBL and 0 when learning through LD. Slightly more students reached a higher level of explanation when learning took place through PBL-individual compared to LD (*M*s = 2.68 vs. 2.26), but this difference did not reach significance, *Z* = –1.11, *p* = .267. Nine students achieved a higher level of explanation when learning via PBL-individual, 4 students achieved a higher level of explanation when learning via LD, and 6 students achieved the same level of explanation in each.

PBL-individual versus LD application. For the application assessment, mean percentage of concepts applied when learning via PBL-individual was 34.59% (*SD* = 27.07; range 0%–100%), and mean percentage of concepts applied when learning via LD was 12.5% (*SD* = 16.67; range 0%–50%). This difference was significant, *Z* = –3.31, *p* = .001. Fifteen students applied more concepts when learning via PBL-individual, only 1 student applied more concepts when learning via LD, and 3 students applied the same number of concepts in both topics. The majority of students also reached a higher level of explanation when learning took place via PBL-individual compared to LD (*M*s = 2.11 vs. 1.16), *Z* = –2.23, *p* = .026. Eleven students achieved a higher level of explanation when learning via PBL-individual, only 2 students achieved a higher level of explanation when learning via LD, and 6 students achieved the same level of explanation in each.

Analyses of individual patterns in Class 3 thus also support between-subject analyses in indicating that the majority of students recalled, comprehended, and applied concepts better when learning took place via PBL compared to LD.

Discussion

The evidence presented here conceivably will be seen by some as showing little more than what is taken as a given on the part of a good number of practitioners: Students show better long-term retention and ability to apply new material if the instructional method is one that actively engages them and enables them to put new ideas to use. Yet we see the present research as more than a rigorous demonstration of the obvious. Rather, we see it as a starting point, not a final seal of approval, on a path toward the important goal of solidly evidence-based instructional practice. The next steps involve further rigorous investigation of the mechanisms by means of which a promising instructional method achieves its effects.

A major strength of the research reported here is that it was conducted with a high degree of experimental control within a classroom setting. Students learned in their familiar, real-world school setting, enhancing external validity, while the experimental design maximized internal validity, with instructor, curricular objectives, content, and schedule of instruction equated across conditions. A further factor critical to internal validity is the fact that this intervention was shorter than those typical in the PBL literature, many of which span a full year, and was targeted to very specific and precisely defined learning objectives. With longer interventions and broader, more extensive learning objectives, it is much more difficult to measure outcomes precisely, as well as to thoroughly control for extraneous variables confounding treatment, such as variations in time and exposure to treatment, differing instructors with teaching styles, and varying student participation (Albanese & Mitchell, 1993; Vernon & Blake, 1993). Thus, a short, targeted intervention made it possible to minimize confounding variables and to put the different methods of instruction to a rigorous test while nonetheless maintaining the external validity of a naturalistic setting.

The narrowly targeted learning objectives and the crossed design also greatly reduce, even if they do not eliminate, concern regarding experimenter bias. The researcher is not simply favoring (consciously or unconsciously) one group of students (or one topic). We made every effort to instantiate "best practice" versions of PBL and LD instructional practice. Still, we cannot rule out the possibility that the instructor consciously or unconsciously delivered a superior product in one case due to subtle differences that were extraneous to the definitions of each practice. In the case of PBL practice, this potential influence is diluted by the presence of multiple coaches and the indirect role they play in instruction. As emphasized earlier, PBL is not an unstructured method, but the structure lies more in the design of the instruction than in its delivery. Given the present results, it could be worthwhile to undertake further research in which instruction is delivered by instructors who are indifferent and uninformed about the instructional methods (beyond the minimum instruction necessary to provide them regarding what they are to do in the classroom). This strategy, however, might yield poor instantiations of all methods. A more promising strategy might be to engage expert observers to study videotapes of instructional delivery, examining multiple factors related to both intellectual and affective quality of instructors' communications to students.

Would the present findings generalize to instruction of greater length and complexity? Again, this is a question that only further research can answer with assurance. The challenge in undertaking such research, however, will be the inverse relation that exists between duration of instruction and possibility for experimental control. The design of the present study could not readily be replicated with instruction of semester-long duration without introducing confounding factors. In our view, the best strategy will be to very gradually increase length of instruction (as well as to vary other factors such as subject matter and grade level) to ascertain how broadly the present findings extend.

Our PBL groups were staffed by two extra coaches, in addition to the primary teacher. Because employing extra coaches is not always feasible, further research is needed to explore how well our findings apply to PBL classes utilizing other types of scaffolds as well as classes with less able student populations. Other topics for future research with respect to the challenges of implementing PBL with pre-college students include designing effective, developmentally appropriate problems; determining how much scaffolding is necessary for different populations; and training teachers in their roles as PBL coaches. Like those for any kind of instruction, PBL outcomes depend heavily on the skill of those who implement it.

We turn now to what is perhaps the most fundamental question: Why did PBL instruction yield superior comprehension and application of new material? One thing the present results allow us to say definitively is that this superiority cannot be attributed to the social component of PBL. The between-group data showed a leaning toward PBL-team superiority and the within-group data a leaning toward PBL-individual superiority, but the differences in neither case were significant. The comparisons of PBL and LD conditions showed PBL to be more effective in fostering both comprehension and application of concepts—with the superiority for comprehension especially striking since providing definitions for concepts is the type of task that traditional instruction is expected to support well.

Also striking is the consistency of these results with results for adult students in studies of parallel design by Capon and Kuhn (2004) and Pease and Kuhn (2011), as well as with several recent meta-analyses (Strobel & van Barneveld, 2009; Walker & Leary, 2009). Although PBL performance was superior to LD, performance under PBL-team and PBL-individual conditions did not differ significantly. Social collaboration has been included as a standard feature of PBL in applied settings. It is thought to reduce the burden of learning, especially for novice learners who lack relevant knowledge and skills, by distributing the cognitive load (Schmidt et al., 2007). However, contrary to popular thinking about PBL, and consistent with findings by Pease and Kuhn (2011) for college students, the present research indicated that collaboration is not one of the essential components of PBL.

In undertaking to experimentally subtract the social component from PBL, and in highlighting these decisive results, we in no way wish to promote a more general conclusion that collaborative educational methods yield no benefit. Clearly any comparison of instructional methods will depend on the particularities of the instantiations of each method. What we can tentatively conclude is that the effectiveness of PBL cannot be attributed exclusively to the benefit that students accrue from engaging with one another. Social interaction by itself is not a "magic bullet" that benefits students. A fundamental question is who in the group benefits? It has been found in some cases that the superior performance of the group is due simply to the increased probability of someone in the group possessing the needed expertise (Kuhn, Pease, & Wirkala, 2009). Collaborative learning situations demand the same painstaking analysis of learning as do all learning situations, made more complex by the fact that multiple learners are involved and their learning affects that of their peers in complex ways. Until more such microgenetic observations of collaborative learning are carried out, we are limited in what we can conclude about its nature.

Quite distinct from this social dimension, PBL appears to provide benefits attributable to its defining core, engagement with problems. Certain of its characteristics appear to be particularly conducive to pre-college students' learning. First, the problem provides a potentially motivating, goal-based activity, particularly important for young students. Although we lack independent evidence that students were more motivated in PBL conditions, there are reasons to believe that motivation was a contributing factor. Compared to adults in the Capon and Kuhn (2004) and Pease and Kuhn (2011) PBL studies, the middle school students in the present work not only had much lower levels of preexisting knowledge and expertise, but because they were not engaged in a course of study of their own choosing, it is likely that they had lower intrinsic motivation than adult students. Thus, the sequence of goal-oriented, inquiry-like activities (asking questions, identifying learning gaps, finding evidence, revising explanations, etc.) may serve as an effective scaffold that heightens young students' cognitive and affective engagement. Moreover, the problems are authentic; students recognize them as important and worth thinking about and recognize that they could apply to their lives outside of school, all of which may increase motivation. Finally, the problem provides context or a "storyline" that new concepts can fit into, especially important for novice learners. Although these factors may have contributed to PBL students' learning, we have no independent evidence of their role in accounting for the differences observed across groups. Establishing the role of each of them remains a task for further research.

Lecture, in contrast, is thought to be effective in part because it provides students with multiple examples that reinforce their learning (Van Merrienboer & Sweller, 2005). However, without a coherent and memorable context, the information may not be as effectively encoded and stored in memory. The power of context was seen in the application assessment, in which multiple students specifically referred back to the PBL problem ("this is just like the NASA problem . . . ") while explaining the concepts' application to the new situation. Thus, although LD students in this study arguably should have learned more because they received more information, there is no benefit if the student is unable to recall, comprehend, or apply the information.

Activation of prior knowledge is another characteristic of PBL that likely contributes to its effectiveness. PBL students engage with the problem prior to accessing information that would help them solve it. Because they cannot simply rely on information given to them, they have no recourse but to access their prior knowledge. Thus, existing knowledge structures get activated, although not necessarily through discussion with others, as sometimes assumed (Norman & Schmidt, 1992; Schmidt, De Grave, De Volder, Moust, & Patel, 1989; Schmidt et al., 2007), since students working alone did as well as those who conferred with peers. The initial retrieval process, during which students struggle with a problem and access prior knowledge they believe relevant, may lead PBL students to activate more retrieval paths that will connect to the new concepts they learn and apply. New ideas can be encoded into the activated prior knowledge structures. Thus, association with older, familiar information makes new knowledge more *meaningful*, and because there are more retrieval pathways, it also becomes more *accessible*. This characterization is consistent with Schwartz and Bransford's (1998) view that there is indeed a "time for telling," but that time should be after students have first struggled to make sense on their own.

Finally, PBL entails generating explanations and elaborations, which include making inferences about new knowledge in order to connect it and apply it to the problem. Thus, the metastrategic competencies that are involved in consciously reflecting on one's own thinking and revising it in the light of new evidence may be strengthened as an outcome of PBL (Kuhn, 2005; Kuhn & Pease, 2006, 2008, 2009). And as Bransford, Brown, and Cocking (2000) describe, teaching practices congruent with a metacognitive approach to learning, which are focused on sense-making and reflection on one's learning, may increase the degree to which students can transfer their learning to new settings and events. The 9-week delay between instruction and assessment in the design of our study made it possible for outcome measures to assess this deep-level learning.

Do the benefits of PBL extend beyond mastery of targeted knowledge and understanding to include learning skills and dispositions themselves? Especially for the K–12 population, this is a critical question. It is one, however, that can only be answered with continuing rigorous investigations of both processes and products of PBL, including investigation of students' own assessments of their learning experience. The cognitive skills and dispositions that PBL might foster warrant analysis in their own right, as does experimental analysis of the still multicomponent process that PBL consists of. Although the present work focuses on outcomes rather than process, we believe our findings indicate that the more laborious process observations and analyses warrant the investment.

Notes

[1] The "cognitive diversity"concept was split into two concepts ("cognitive" diversity and "ideas"), thus creating an 8th concept for this topic. This was useful due to the richness and wide variety of accurate applications of this concept.

[2] The choice of two topics unrelated to one another minimizes the possibility of order effects. Comparison of Class 2 (problem-based learning [PBL], then lecture/discussion [LD]) and Class 3 (LD, then PBL) addresses the possibility of an order effect between PBL and LD. The order of the two PBL conditions was not varied, but performance did not differ across conditions. The possibility remains that a reverse order (team first) could have produced different results. However, it is the order we used (individual, then team) that theory would predict most likely to manifest a PBL-team superiority, and this did not appear.

References

Albanese, M., & Mitchell, S. (1993). Problem-based learning: A review of literature on its outcomes and implementation issues. *Academic Medicine*, *68*, 52–81.

Berkson, L. (1993). Problem-based learning: Have the expectations been met? *Academic Medicine*, *68*(Suppl. 10), S79–S88.

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

Capon, N., & Kuhn, D. (2004). What's so good about problem-based learning? *Cognition and Instruction*, *22*(1), 61–79.

Colliver, J. (2000). Effectiveness of problem-based learning curricula: Research and theory. *Academic Medicine*, *75*, 259–266.

Davies, P. (2000). Approaches to evidence-based teaching. *Medical Teacher*, *22*(1), 14–21.

Dochy, F., Segers, M., Van den Bossche, P., & Gijbels, D. (2003). Effects of problem-based learning: A meta-analysis. *Learning and Instruction*, *13*, 533–568.

Gijbels, D., Dochy, F., Van den Bossche, P., & Segers, M. (2005). Effects of problem-based learning: A meta-analysis from the angle of assessment. *Review of Educational Research*, *71*(1), 27–61.

Higbee, K. L. (2001). *Your memory: How it works and how to improve it* (2nd ed.). New York, NY: Marlowe & Company.

Hmelo-Silver, C. (2004). Problem-based learning: What and how do students learn? *Educational Psychology Review*, *16*(3), 235–266.

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

Kirschner, P., Sweller, J., & Clark, R. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*(2), 75–86.

Kuhn, D. (2005). *Education for thinking*. Cambridge MA: Harvard University Press.

Kuhn, D., & Pease, M. (2006). Do children and adults learn differently? *Journal of Cognition and Development*, *7*, 279–293.

Kuhn, D., & Pease, M. (2008). What needs to develop in the development of inquiry skills? *Cognition and Instruction*, *26*, 512–559.

Kuhn, D., & Pease, M. (2009). The dual components of developing strategy use: Production and inhibition. In H. S. Waters & W. Schneider (Eds.), *Metacognition, strategy use, and instruction* (pp. 135–159). New York, NY: Guilford.

Kuhn, D., Pease, M., & Wirkala, C. (2009). Coordinating effects of multiple variables: A skill fundamental to causal and scientific reasoning. *Journal of Experimental Child Psychology*, *103*, 268–284.

Lahey, B. B. (2008). *Psychology: An introduction* (10th ed.). New York, NY: McGraw-Hill.

Lorayne, H., & Lucas, J. (1974). *The memory book: The classic guide to improving your memory at work, at school, and at play*. New York, NY: Ballantine Books.

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? *American Psychologist*, *59*, 14–19.

Mitchell, K., Shkolnik, J., Song, M., Uekawa, K., Murphy, R., Garet, M., & Means, B. (2005). *Rigor, relevance, and results: The quality of teacher assignments and student work in new and conventional high schools*. Washington, DC: American Institutes for Research and SRI International.

Norman, G., & Schmidt, H. (1992). The psychological basis of problem-based learning: A review of evidence. *Academic Medicine*, *67*, 557–565.

Pease, M., & Kuhn, D. (2011). Experimental analysis of the effective components of problem-based learning. *Science Education*, *95*, 57–86.

Schmidt, H. G. (1993). Foundations of problem-based learning: Some explanatory notes. *Medical Education*, *27*, 422–432.

Schmidt, H. G., De Grave, W. S., De Volder, M. L., Moust, J. H. C., & Patel, V. L. (1989). Explanatory models in the processing of science text: The role of prior knowledge activation through small-group discussion. *Journal of Educational Psychology*, *81*, 610–619.

Schmidt, H. G., Loyens, S. M. M., van Gog, T., & Paas, F. (2007). Problem-based learning is compatible with human cognitive architecture: Commentary on Kirschner, Sweller and Clark (2006). *Educational Psychologist*, *42*(2), 91–97.

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

Schwartz, D., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction*, *22*(2), 129–184.

Strobel, J., & van Barneveld, A. (2009). When is PBL more effective? A meta-synthesis of meta-analyses comparing PBL to conventional classrooms. *The Interdisciplinary Journal of Problem-Based Learning*, *3*(1), 44–58.

Surowiecki, R. (2005). *The wisdom of crowds*. New York, NY: Anchor Books.

Van Merrienboer, J. J. G., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review*, *17*(2), 147–177.

Vernon, D. T. A., & Blake, R. L. (1993). Does problem-based learning work? A meta-analysis of evaluative research. *Academic Medicine*, *68*, 550–563.

Walker, A., & Leary, H. (2009). A problem based learning meta analysis: Differences across problem types, implementation types, disciplines, and assessment levels. *The Interdisciplinary Journal of Problem-Based Learning*, *3*(1), 12–43.

Manuscript received September 4, 2010
Final revision received May 13, 2011
Accepted May 18, 2011