Journal of Educational Psychology

Comparing Learning From Observing and From Human Tutoring

Kasia Muldner, Rachel Lam, and Michelene T. H. Chi

Online First Publication, October 14, 2013. doi: 10.1037/a0034448

CITATION

Muldner, K., Lam, R., & Chi, M. T. H. (2013, October 14). Comparing Learning From Observing and From Human Tutoring. *Journal of Educational Psychology*. Advance online publication. doi: 10.1037/a0034448

Journal of Educational Psychology

© 2013 American Psychological Association
0022-0663/13/$12.00 DOI: 10.1037/a0034448

Comparing Learning From Observing and From Human Tutoring

Kasia Muldner, Rachel Lam, and Michelene T. H. Chi
Arizona State University

A promising instructional approach corresponds to *learning by observing others learn* (i.e., by watching tutorial dialogue between a tutor and tutee). However, more work is needed to understand this approach's pedagogical utility. Thus, in 2 experiments we compared student learning from collaborative observation of dialogue with 2 other instructional contexts: 1-on-1 human tutoring and collaborative observation of monologue. In Study 1 ($N = 50$), there was no significant difference in learning outcomes between the dialogue and tutoring conditions, while the dialogue condition was superior to the monologue condition. Study 2 ($N = 40$), which involved a younger population than in Study 1, did not replicate these results, in that students learned less from observing dialogue than from being tutored, and there was no significant difference between the dialogue and monologue conditions. To shed light on our results, we analyzed the verbal data collected during the 2 experiments. This analysis showed that in Study 1, the dialogue observers generated more substantive contributions than did the monologue observers. In contrast, in Study 2 there was no significant difference between the observers in terms of substantive contributions; moreover, the total number of contributions was modest, which may have hindered observer learning in that study. In general, our findings suggest that collaboratively observing tutorial dialogue is a promising learning paradigm, but more work is needed to understand how to support young students to effectively learn in this paradigm.

Keywords: collaborative observation, human tutoring, emergent topics

Student learning can take place in a variety of instructional contexts, including studying alone, one-on-one tutoring, and collaborative group activities with the aid of course textbooks, to name a few. A less conventional but highly promising instructional approach corresponds to *learning by observing others learn* (i.e., by watching tutorial dialogue between a tutor and tutee). However, more work is needed to understand this approach's pedagogical utility and, in particular, how it compares to other instructional contexts. Through two experiments, we take steps to fill this gap by comparing student learning from collaborative observation of tutorial dialogue to two other forms of instruction, namely, one-on-one human tutoring and collaborative observation of tutorial monologue.

Learning From Tutoring and From Observing

Many studies have demonstrated that one-on-one human tutoring is an effective strategy for fostering student learning (Bloom, 1984; P. A. Cohen, Kulik, & Kulik, 1982; Graesser, Person, & Magliano, 1995; Lepper, Woolverton, Mumme, & Gurtner, 1993; Merrill, Reiser, Merrill, & Landes, 1995). For instance, a recent metareview showed that compared to nontutoring contexts like standard classroom instruction, the effect size of one-on-one human tutoring was high at $d = 0.79$ (VanLehn, 2011). However, providing a tutor for every student is not feasible, and so there have been various efforts to identify other more scalable contexts that afford students the benefits of individualized instruction. One area of research focuses on developing intelligent tutoring systems (ITSs), which are computer-based applications that aim to provide personalized instruction by adapting to a given student's needs. While there is evidence that ITSs can benefit student learning (Arroyo, Beal, Murray, Walles, & Woolf, 2004; Koedinger, Anderson, Hadley, & Mark, 1995; VanLehn et al., 2005), and in some cases even match human tutors' effectiveness (VanLehn, 2011), these systems suffer from two key limitations. First, they require substantial development time—estimates range from 100 to 1000 hr of development per hour of instruction (Anderson, 1993; Murray, 1999). Second, the majority of ITSs target procedural domains and skills (VanLehn et al., 2005; Wang & Heffernan, 2011), and so less support is available for open-ended conceptual domains and domain-independent skills like collaboration or self-explanation (with notable exceptions; e.g., Aleven, McLaren, Roll, & Koedinger, 2006; Muldner & Conati, 2010; Walker, Rummel, & Koedinger, 2011).

To address the limitations of ITSs but simultaneously take advantage of the benefits of tutoring, Chi, Roy, and Hausmann (2008) proposed an alternative instructional context, which they referred to as *learning from observing others learn*. As the name suggests, this context corresponds to situations where students acquire complex cognitive skills by observing (and overhearing)

Kasia Muldner, Department of Psychology, Arizona State University; Rachel Lam, Mary Lou Fulton Teachers College, Arizona State University; Michelene T. H. Chi, Learning Sciences Institute, Mary Lou Fulton Teachers College, Arizona State University.

This research was supported by a grant from the Spencer Foundation (200800196). We thank Ruth Wylie and Erin Walker for their helpful comments on earlier drafts of the article, Samuel Green for statistical advice, and Omid Vasefi for assisting with the preparation of the study materials and help in conducting the studies.

Correspondence concerning this article should be addressed to Kasia Muldner, who is now at the School of Computing, Informatics and Decision Systems Engineering, Arizona State University, 699 South Mill Avenue, Brickyard, Tempe, AZ 85281. E-mail: Katarzyna.Muldner@asu.edu

others learn, such as learning to solve physics problems by observing a video of a tutor helping a student solve such problems at a whiteboard. Thus, the learning from observing others learn paradigm differs from traditional work that has focused on how individuals learn overt behaviors, also referred to as vicarious learning (e.g., children learning to act aggressively by watching aggressive behaviors; Bandura, Ross, & Ross, 1961). Given that learning from observing others learn involves a dialogue between a tutee and a tutor as they go over instructional materials, here we refer to this paradigm as *observing dialogue.*

The observing dialogue paradigm can be implemented by creating instructional videos that are easily disseminated and shared, making this approach relatively low cost and scalable. However, prior to the Chi et al. (2008) study, earlier work suggested that tutoring (or interactions resembling it) generated superior performance compared to observing. For instance, Schober and Clark (1989) had participants serve in a tutee role to solve Tangram puzzles with the direct guidance of an experimenter (acting in a tutor role). Other participants (referred to as observers) solved the same puzzles alone while overhearing the interactions of the tutee and tutor but not directly participating in their dialogue. The results showed that the tutees' performance was superior to the observers' performance. Craig, Driscoll, and Gholson (2004, Experiment 1) also found that students who merely watched dialogue videos about computer literacy topics learned less than students who interacted with a computer tutor.

In both the Schober and Clark (1989) study and the Craig et al. (2004, Experiment 1) study, the tutees could interact directly with a tutor, while the observers could only passively watch these interactions. Thus, the tutees had an advantage, given that the benefits of interaction are demonstrated by various studies (e.g., E. G. Cohen, 1994; Johnson & Johnson, 2009) and conceptual frameworks (Chi, 2009). For instance, the interactive-constructive-active-passive (ICAP) framework (Chi, 2009), which differentiates learning activities according to overt student behaviors, makes the prediction that a student who interacts constructively with another individual will, in general, learn better than a student working alone. This prediction is based on the fact that interaction offers students the opportunity to explain their perspective, elicit responses from a partner, and integrate a partner's contributions, to name a few. Given these various benefits of interaction, it is therefore not surprising that in the aforementioned studies the tutees performed better than the observers, but it is not clear whether this was due to the presence of a tutor or because the observers were passive rather than constructively interactive.

To address this issue, in their work comparing learning from being tutored and from observing others learn, Chi et al. (2008) had some of the observers work in pairs, thus providing them with interaction opportunities. The conjecture was made that if observers have a partner to work with and a tutorial dialogue video to observe, their learning would be substantially improved. To test this conjecture, some students (the tutees) were videotaped solving procedural physics problems while interacting with an expert tutor. Other students, working either in pairs or alone, either watched the videos of these tutorial dialogue sessions while solving the same physics problems as the tutees or used a textbook while solving the problems. Although all conditions showed gains from pretest to posttest, no significant difference was found between the collaborative observers watching the tutorial dialogue videos and the tutees in the videos. In contrast, students who observed alone and dyads who used the textbook learned significantly less than the tutees. This suggests that if observers interact with a peer, then dialogue videos are effective instructional materials.

Other work has focused on comparing outcomes from observation of a tutorial dialogue versus a tutorial monologue, the latter showing a tutor going over the instructional materials alone without a tutee, much as a teacher would in a classroom lecture (Cox, McKendree, Tobin, Lee, & Mayes, 1999; Craig, Chi, & VanLehn, 2009; Craig, Gholson, Ventura, & Graesser, 2000; Craig, Sullins, Witherspoon, & Gholson, 2006; Driscoll, Craig, Gholson, Hu, & Graesser, 2003; Fox Tree, 1999; Fox Tree & Mayer, 2008; Muller, Bewes, Sharma, & Reimann, 2008; Muller, Sharma, Eklund, & Reimann, 2007; Schunk & Hanson, 1985). The majority of these studies show that students learn more from observing dialogue than monologue. Various reasons have been proposed for the benefits of dialogue observation (for a review, see Chi, 2013), such as that compared to a monologue, a dialogue encourages collaborative observers to be more engaged (Craig et al., 2009), as well as that dialogue includes beneficial features like (1) misconceptions generated by a tutee and refuted by a tutor (Muller et al., 2007; Schunk & Hanson, 1985; Schunk, Hanson, & Cox, 1987) and (2) "deep" questions, ones that require more than a one word answer, posed by the tutor (Craig et al., 2006, Experiment 1; Driscoll et al., 2003).

Given that the only study to compare learning from observing dialogue and human tutoring did not find a significant difference (Chi et al., 2008) and that other studies show observing dialogue to be more effective than observing monologue, this suggests that there is something unique about dialogue and requires replication of both conditions in a single study, as we do here.

Present Studies

We report on two studies analyzing student learning from observing and being tutored, with two target comparisons. Our first target comparison corresponds to *collaboratively observing dialogue* versus *human tutoring*. To the best of our knowledge there is only one study comparing learning from these two instructional contexts (Chi et al., 2008). While the researchers in that study did not find a significant difference between the dialogue and tutoring conditions, studies comparing being tutored to other activities have predominantly and consistently shown tutoring to produce better learning (e.g., P. A. Cohen et al., 1982). Thus, clearly more work is needed to understand the relative benefits of tutoring and observing tutorial dialogue.

Our second target comparison corresponds to *collaboratively observing dialogue* versus *collaboratively observing monologue*. As indicated above, there is evidence that students learn more from observing dialogue videos (2009, Craig et al., 2000, 2006; Driscoll et al., 2003; Fox Tree, 1999; Muller et al., 2007; Schunk & Hanson, 1985). However, the vast majority of this work has scripted the content of the dialogue and monologue videos, manipulating the presence or absence of features of interest. While this work has provided valuable insight into some of the aspects that influence observer learning from dialogue and monologue, scripting has several drawbacks. First, it is an expensive intervention to implement, as designing video content requires an in-depth understanding of various aspects, such as the target domain, common student misconceptions for that domain, common student questions, and appropriate tutor strategies such as scaffolding and prompting. Second, more work is needed to fully understand how to properly script an instructional video, since there is not yet a full picture of all the factors that influence observer learning or the interplay between those factors. In contrast, taping and reusing existing tutorial sessions, for instance ones obtained in a tutoring center, alleviates these scripting challenges.

As far as we are aware, there is only one educational study that relied on unscripted content when comparing dialogue to monologue (Craig et al., 2009); this study involved university students and a physics domain. While the results showed dialogue to be superior to monologue in terms of student performance on a delayed problem-solving task, research is needed to see if the results involving unscripted content generalize to other domains and populations.

Given the considerations above, in our work we used unscripted content when creating the dialogue and monologue videos. Another important property of the videos is the tutors and tutees who appear in them. Prior work used the same tutor in all the videos (Chi et al., 2008; Craig et al., 2009) or produced a single video per condition, meaning that the observers all saw the same tutor (and tutee in the dialogue condition; e.g., Driscoll et al., 2003; Muller et al., 2008). With such designs, it is difficult to determine whether the outcomes related to observer learning are due to a particular manipulation or to a specific tutor or tutee in the video. Thus, we recruited multiple tutors and tutees for our instructional videos, thereby ensuring that our results would not be tied to an individual tutor or tutee and so increasing the generalizability of our findings.

Target Domain

Our target domain corresponded to a conceptual science topic—molecular diffusion. Evidence has suggested that this is a highly misconceived and challenging topic for students (Chi, 2005; Chi, Roscoe, Slotta, Roy, & Chase, 2012; Meir, Perry, Stal, & Klopfer, 2005), because it requires understanding of two difficult concepts: emergent processes and proportionality.

In general, an emergent process is one that includes many micro-level agents that behave according to simple rules to produce a more complex, macro-level pattern or outcome (Levy & Wilensky, 2008). For the case of molecular diffusion, molecules (micro-level agents) behave according to the rule of continuous random motion. The macro-level pattern arising from these molecular interactions can be perceived as a flow of one substance into another prior to equilibrium or as a stable, unchanging solution at equilibrium. Chi (2005) and Chi, Roscoe, et al. (2012) proposed that molecular diffusion can be regarded as a decentralized system, which is a general characteristic of emergent processes (Resnick, 1996), since there is no controlling agent directing the behavior of molecules. Moreover, Chi (2005) provided a precise characterization of an emergent process, by specifying a set of 10 *features* and *attributes* that identify and explain such a process (where features characterize micro-level aspects and attributes characterize inter-level connections between the micro and macro levels). Our instructional materials targeted these features and attributes.

To illustrate, suppose some blue dye is dropped into water. The subsequent diffusion of dye throughout the water is an emergent process for the following reasons:

- (random feature) The molecular interactions are random, in that any molecule can collide or interact with any other molecule.
- (disjoint attribute) The dye and water molecules and the visible flow pattern of the dye can behave in disjoint ways. For instance, the flow (macro pattern) may appear to move in a certain direction, while the molecules (micro agents) are bouncing around and colliding in ways that may go against this macro pattern.
- (collective attribute) The flow pattern of dye is caused by the collective summing of all the molecular interactions.

For the full list of emergent features and attributes, see Chi, Roscoe, et al. (2012). From the perspective of learning, the two emergent attributes listed above are classified as inter-level because they require students to reason about both the visible macro-level pattern and the underlying micro-level interactions. Students hold various inter-level misconceptions, such as that the molecules stop moving at equilibrium because the solution appears to be a uniform, unchanging color (Meir et al., 2005). This misconception suggests that students think the pattern at the micro-level must correspond to the pattern at the macro level, thus contradicting the "disjoint" attribute. Students also hold misconceptions about micro-level features, such as believing that dye molecules aim to move to areas in the solution where there are fewer dye molecules and thus "more room," instead of by spreading through random collisions.

The collective attribute requires understanding of ratio and proportion. For instance, in the context of the above dye-in-water example, the changes in concentration of dye relative to water from one area of the beaker to another is what allows one to see the visible flow of the dye throughout the water. Numerous studies have shown that proportion concepts are difficult for students (e.g., Smith, Carey, & Wiser, 1985), further adding to the complexity of learning about molecular diffusion.

In general, learning about diffusion is challenging because students hold many preconceptions that contradict scientifically appropriate notions for emergent phenomena, as described in detail in Chi, Roscoe, et al. (2012). For instance, the belief that all processes follow a sequential, linear progression is reinforced in children because there are many examples of sequential processes in nature, but this characteristic does not hold for emergent processes. Thus, when students are exposed to emergent processes like diffusion, they develop misconceptions that result from their prior knowledge. Here we explore student learning about the emergent process of diffusion in three instructional contexts, and in particular the pedagogical utility of observing tutorial dialogue for this domain. As described above, a tutorial dialogue encourages student interaction (Craig et al., 2009), which in turn may foster learning. Moreover, a dialogue includes beneficial features like tutee misconceptions and questions. For instance, since students hold the misconception that molecules become still at equilibrium, observers of a dialogue video may benefit from witnessing a tutor refute this misconception after a tutee expresses it.

Study 1

In Study 1, we used the emergent conceptual domain described above and university-level participants to compare student learning from (1) observing tutorial dialogue versus one-on-one human tutoring and (2) observing tutorial dialogue versus observing tutorial monologue. In the tutoring condition, each student worked individually with a human tutor to solve a set of problems. To provide students in all conditions with interaction opportunities (the observers as well as the tutees), in the dialogue and monologue conditions students worked in pairs to solve the same set of problems as were solved by the tutees.

Methodology

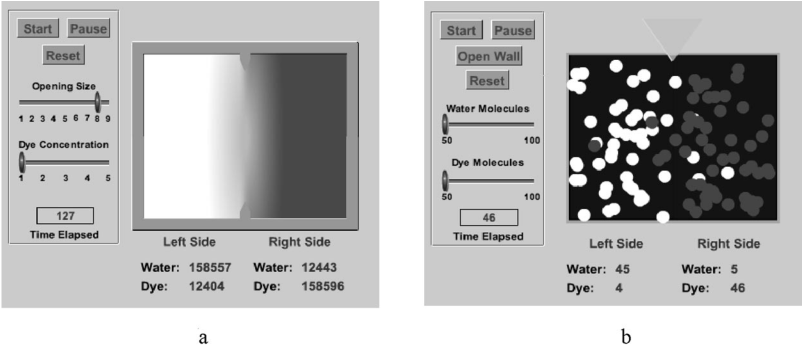
Materials. The study involved the following materials related to diffusion: (1) a two-page diffusion text, (2) a diffusion pretest and posttest, (3) two diffusion simulations, (4) a diffusion workbook, and (5) 20 instructional videos (10 dialogue and 10 monologue).

Diffusion text + tests. The diffusion text was designed to provide a general overview of diffusion to help prepare students for study participation. The pretest and posttest assessed students' diffusion knowledge (the pretest included 25 multiple-choice questions, while the posttest included the same 25 questions and four other questions, for a total of 29 questions). The tests included questions that probed understanding of emergent aspects of diffusion but without explicitly mentioning emergence. For instance, to assess knowledge of the inter-level disjoint attribute, one question asked "As the dye diffuses away from where it was originally dropped into the water, can some dye molecules bounce back towards this original place?" (see Appendix for more examples).

Diffusion simulations. To help students understand diffusion concepts, the two simulations showed diffusion occurring at the visible level (*macro simulation;* see Figure 1a) and at the molecular level (*micro simulation;* see Figure 1b). For instance, clicking the *start* button in the micro simulation results in molecules bouncing and colliding in their container. However, participants could only observe the simulations being played and could not manipulate them, consistent with the methodology used in a prior study (Chi, Kristensen, & Roscoe, 2012). This decision was based on repeated observations from pilot evaluations showing that when students could interact with this type of simulation, they did not use it effectively (e.g., failed to use it at all or manipulated irrelevant features; Muldner, Dybvig, Lam, & Chi, 2011). Thus, to avoid ineffective use of the simulations, in the present experiment the tutor controlled the simulations when showing them to the tutees.

Diffusion workbook. To help guide students' activities during the experiment, we created a diffusion workbook that students were asked to complete. The workbook consisted of seven problems, one per page (each with two to four questions), and covered the following topics: concentration, proportionality, flow, molecular interactions and behavior, diffusion of liquids and gases, and diffusion across a semipermeable membrane. As was the case for the diffusion tests, the workbook problems were designed to indirectly address emergent features and attributes without explicitly referring to emergence. The majority of these problems included various surface differences with the test questions to avoid a teaching-to-the-test effect.

To answer the workbook problems, students had to draw and write. For instance, Figure 2 shows one of the workbook problems, consisting of three questions based on a test item in Meir et al. (2005). The problem describes how when dye is poured into the left side of a rectangular container of water, it flows from where it was dropped toward the other areas of the container and, in particular, appears to flow to the right (this is the macro-level, or visible view, description). For Question 1, students were provided with a micro-level (molecular) diagram that contained the dye and water molecules corresponding to the macro-level description and were asked to draw the direction(s) that a given *individual* molecule was likely to move. The correct solution reflects that for an individual molecule, any direction is just as likely, since molecules move randomly. Thus, this activity targets the disjoint attribute listed above, because molecules can move in directions opposite to the visible flow pattern. However, students may incorrectly believe that an individual molecule is more likely to move in the direction of the dye flow, from higher to lower concentration. The other two questions (see Figure 2) further probed students to consider the relation between the macro-level flow of dye and the underlying dye and water molecular interactions, targeting their understanding of the random feature and disjoint attribute, respectively.



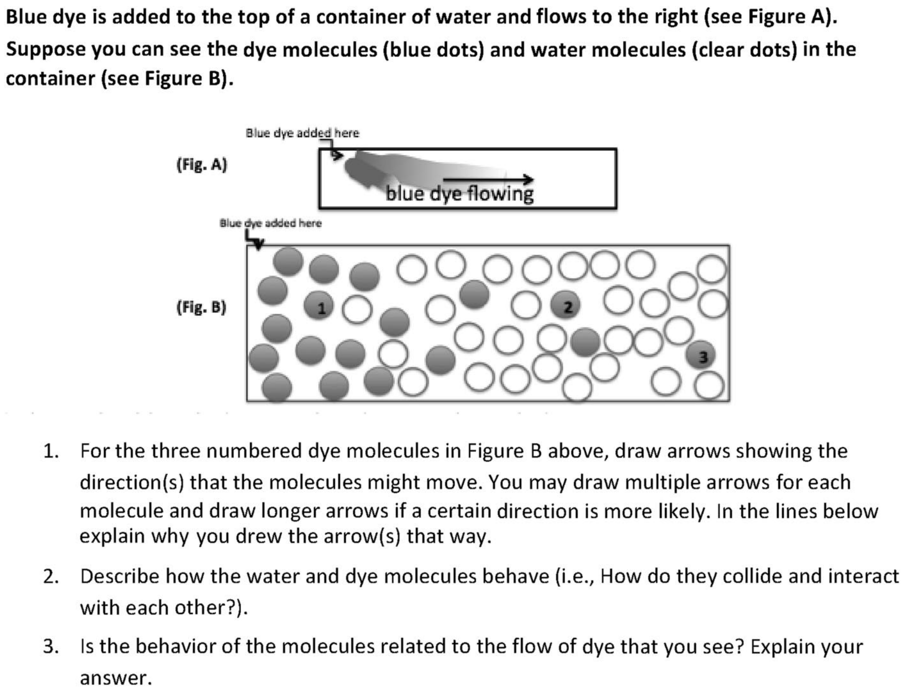
127 Time Elapsed

Left Side Water: 158557 Dye: 12404
Right Side Water: 12443 Dye: 158596

46 Time Elapsed

Left Side Water: 45 Dye: 4
Right Side Water: 5 Dye: 46

Figure 1. The two simulations showing diffusion (a) at the visible macro level and (b) at the molecular micro level.



Blue dye is added to the top of a container of water and flows to the right (see Figure A). Suppose you can see the dye molecules (blue dots) and water molecules (clear dots) in the container (see Figure B).

Blue dye added here
(Fig. A)
blue dye flowing
Blue dye added here
(Fig. B)
1
2
3

1. For the three numbered dye molecules in Figure B above, draw arrows showing the direction(s) that the molecules might move. You may draw multiple arrows for each molecule and draw longer arrows if a certain direction is more likely. In the lines below explain why you drew the arrow(s) that way.
2. Describe how the water and dye molecules behave (i.e., How do they collide and interact with each other?).
3. Is the behavior of the molecules related to the flow of dye that you see? Explain your answer.

Figure 2. Sample workbook questions.

Instructional videos. The instructional videos used in the experiment's observing conditions corresponded to unscripted tutorial dialogue and lecture-style monologue sessions. Neither the dialogue nor the monologue videos were edited, and so the observers saw the sessions exactly as they occurred.

We created each dialogue video by videotaping a tutor helping a student answer the diffusion workbook questions (details on the tutors and tutees are below). To ensure that the observers could clearly see the tutee's work when watching the video, each of the workbook problems was enlarged onto a single laminated poster that was attached to a whiteboard. The workbook posters mainly included the problem images, and the tutor read out loud the problem text. The dialogue videos also included the diffusion simulations, which were projected onto the wall next to the whiteboard upon the tutor's request and shown for the duration of a tutor's instruction around them. Thus, the dialogue videos showed the tutor, tutee, workbook problems on the whiteboard, and the projected simulations (see Figure 3 for an example). To keep these sessions as natural as possible, the tutors decided how to cover the instructional materials, including when and how to use the simulations (all tutors used the simulations to various degrees). Both the tutors and tutees were free to draw and write on the whiteboard, ask questions, and so on.

We created a total of 10 dialogue videos in this fashion, by recruiting five tutors (two female) and having each tutor work individually with a male student on one occasion and a female student on another occasion. The average length of a tutoring video was 25 min. We also created a total of 10 monologue videos, using the same five tutors; each tutor was featured in two monologue videos created shortly before or after his/her two dialogue sessions. For each monologue video, a tutor used a lecture-style presentation to go over the diffusion workbook problems at the whiteboard, also showing the diffusion simulations to illustrate various concepts. As was the case for the dialogue sessions, it was up to the tutors to decide how to cover the instructional materials. These monologue sessions were videotaped in the same fashion as the tutoring sessions. The average length of a monologue video was 21 min.

Since we did not script the content of the dialogue and monologue videos, variations between the videos arose from, for instance, how a tutor chose to explain a concept or a tutee's misconceptions that elicited tutor feedback.[1] However, we did ask the tutors to go over each of the seven workbook problems to ensure that all the videos addressed the key high-level concepts. All tutors predominantly followed this instruction, as we verified by checking the content of the videos (with the following two exceptions: two tutors did not go over one of the worksheet subproblem questions, one during a monologue and one during a dialogue).

Tutor participants and preparation. Two of the five tutors were recruited from the math and science education departments of a large public university in the southwestern United States through flyers and informational meetings. Both tutors had teaching and one-on-one tutoring experience (2+ years). The other three tutors were members of our research lab, including a senior psychology student (with limited tutoring experience), a graduate student in educational psychology (with 10 years' teaching and tutoring experience), and a graduate student in science education (with 2+ years' tutoring experience). Tutors were compensated $30 for participation.

[1] As a side note, tutee or observer learning outcomes were not influenced by which tutor a tutee interacted with or observers watched in a dialogue or monologue video (as we verified with an analysis of covariance with pretest as the covariate and tutor and condition as the independent variables, no significant main effect of tutor or Tutor × Condition interaction was found for either study).

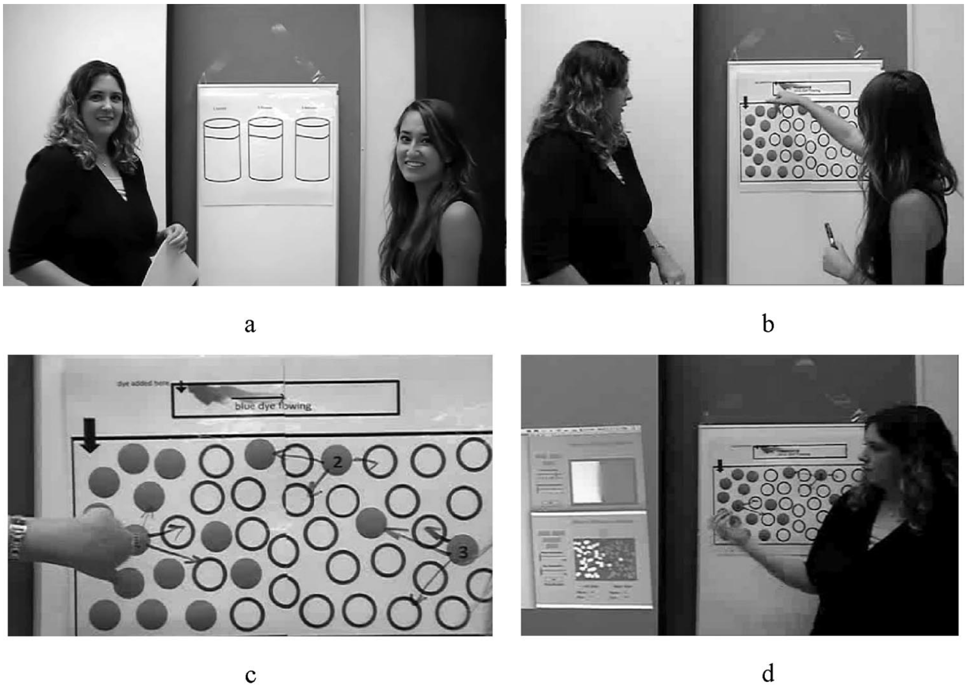


Figure 3. Screenshots of a dialogue video with (a) tutor (left) and tutee (right) getting ready to do the first workbook problem; (b) tutor and tutee discussing the second workbook problem shown in Figure 2; (c) close-up of tutee work as shown in the video; and (d) tutor discussing the diffusion simulations, projected on the left of the whiteboard. Both individuals appearing here gave signed consent for their likenesses to be published in this article.

To prepare the tutors for study participation, we had each tutor (1) read a two-page college-book text on molecular diffusion, (2) read the student diffusion workbook supplemented with solutions, (3) use the two diffusion simulations to complete a "simulation" workbook containing instructions and prompts (e.g., *What does the Macro Simulation look like at equilibrium? What are the dye and water molecules in the Micro Simulation doing during equilibrium?*), and (4) listen to a brief tutorial on pedagogical strategies that differentiated tutor scaffolding, an effective strategy for student learning (Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001), from tutor telling, an ineffective strategy (Chi et al., 2008). In addition, all tutors completed the diffusion posttest and met our criterion of achieving 80% or above on that test.

Student participants. The student participants were university undergraduates ($N = 50$), who completed the study for a first-year psychology course credit, with an equal number of male and female students assigned to each condition.

Design and procedure. Study 1 used a between-subjects design with the following three conditions: (1) collaboratively observing dialogue, (2) one-on-one tutoring, and (3) collaboratively observing monologue.

The procedure for the three conditions was the same. Participants first read the diffusion text (*background phase*) and then completed the diffusion pretest. We borrowed the design of including a background phase *prior* to the pretest from related work (Chi, Bassok, Lewis, Reimann, & Glasser, 1989; Chi et al., 2008). This design captures more accurately the effect of the experimental manipulation than if the background phase is done after the pretest, because the effect of a given intervention is not inflated by including the effect of the background phase. After the pretest, participants completed the diffusion workbook according to each condition's procedure and then filled in the diffusion posttest (the pre- and posttest were done individually). The total time students spent in the study was no more than 2 hr.

For the tutoring condition ($n = 10$, five female), a student worked with a tutor in a private room to answer the workbook questions (details on these sessions are above in the Instructional Videos section). Students were assigned to tutors based on tutor availability, with the constraint that each tutor work with one male and one female student. These sessions were videotaped and transcribed.

For the two observing conditions (dialogue: $n = 20$, 10 female; monologue: $n = 20$, 10 female), students worked in pairs to complete the diffusion workbook while observing a dialogue video (dialogue condition) or a monologue video (monologue condition). Each video was observed by exactly one student pair in a private room. Students viewed the videos on a desktop computer using the VLC media player (www.videolan.org/vlc/) and were given a brief overview of VLC at the start of the session to ensure they could use the player correctly. Because we wanted to avoid gender effects (Harskamp, Ding, & Suhre, 2008), we used same-gender pairs and yoked the observers to the gender of the tutee they saw in the video (e.g., female observers watched female tutees). To encourage the observers to collaborate, each pair was given a single diffusion workbook to share, as advocated by E. G. Cohen (1994).

At the start of a dialogue or a monologue observing session, the participants were told to discuss and answer the workbook questions together and that they could take as long as they needed to do so. Following the methodology in some related work (Chi et al., 2008; Craig et al., 2009, 2004), the observers could control the dialogue and monologue videos by pausing, forwarding, or rewinding, accomplished by using the VLC media player interface tools. Participants were informed about this by being told, *You can rewind/pause/forward the video if you wish—it is up to you as to how you use the video, but we ask that you do use it to help answer the workbook questions and that you pause the video during discussion*. Allowing participants to control the video provided them with opportunities to stop and think about the content (Craig et al., 2004). Moreover, when observers work in pairs, the ability to control the video reduces interference with collaboration opportunities, since students do not have to choose between discussing the target domain and watching the video.

The observers were randomly assigned to the same-gender pairs and to a given dialogue or monologue video from the respective participant pools. After several pairs were run, we began a stratified random sampling procedure based on pretest performance to equalize prior knowledge between the observing conditions. The average length of an observing session was 40 min (38 min and 41 min for monologue and dialogue, respectively, $p = .42$). All sessions were videotaped and transcribed.

Results

Table 1 (top) shows the means and standard deviations for the pretest, posttest, and gain scores in each condition. An analysis of variance (ANOVA) with the pretest data did not reveal significant differences among the conditions, $F(2, 47) = 0.05$, $p = .95$. The average pretest percentage in the three conditions ranged from 55.2% to 57.0% (see Table 1), highlighting that students had some knowledge of diffusion but that this knowledge was quite limited even after reading the diffusion text.

Our primary analysis technique corresponded to analysis of covariance (ANCOVA) with pretest as the covariate, which reduces error variance and is advocated for pretest–posttest designs of the type used in our experiments (Dimitrov & Rumrill, 2003). Planned comparisons were conducted if the ANCOVA found a significant main effect ($p < .05$) or a marginal main effect ($p < .1$); for the latter, some sources have proposed that running planned comparisons is appropriate (Brace, Kemp, & Snelgar, 2003). We calculated effect sizes (eta-squared, η^2) for all reported main effects and Cohen's d for paired planned comparisons, as suggested by Vacha-Haase and Thompson (2004). Cohen's d for the ANCOVA planned comparisons was calculated using pooled variance according to the formula in Howell (2010, p. 610–611). J. Cohen (1988) proposed the following guidelines to interpret effects: as *small* when $\eta^2 = .01$ or $d = 0.2$, *medium* when $\eta^2 = .06$ or $d = 0.5$, and *large* when $\eta^2 = .14$ or $d = 0.8$. However, these guidelines are intended to be general and so are not tailored to a specific discipline. Thus, we also rely on Hattie's (1999) proposal that in the context of educational interventions, an effect size of $d = 0.4$ or greater is considered practically meaningful.

Impact of condition on learning. An ANCOVA with pretest percentage as the covariate and posttest percentage as the dependent variable revealed a significant main effect of condition, $F(2, 46) = 4.23$, $p = .02$, $\eta^2 = .09$; key results for Study 1 are summarized in Table 2 (top). Planned comparisons showed that there was no significant difference in posttest performance (adjusted by the covariate) between the observing dialogue and tutoring conditions (80.9% and 82.6%, respectively), $t(28) = 0.42$, $p = .68$, $d = 0.16$. Note that despite our modest sample size, this comparison should have detected a difference if the effect of being tutored were at the level reported by prior work. In particular, although the only study to compare human tutoring and dialogue observation (Chi et al., 2008) did not report an effect size, the average effect size for being tutored versus other instruction reported in a recent metareview (VanLehn, 2011) was high at $d = 0.79$. This effect size yields a power of 76% (using G*Power software; Faul, 2012), after adjusting for the covariate following the recommendation in Wuensch (2012). However, our sample effect size was much smaller than previously reported at $d = 0.16$, which explains the lack of a significant result (we return to this finding in the General Discussion section). As far as our second target comparison, the paired observers had significantly higher adjusted posttest scores in the observing dialogue condition than in the observing monologue condition (80.9% and 72.6%, respectively), $t(38) = 2.46$, $p = .02$, $d = 0.78$.

Table 1
Means and Standard Deviations for Each Condition at Pretest, Posttest, and Pure Gain in Studies 1 and 2

	Observing dialogue ($n = 20$)		Tutoring ($n = 10$)		Observing monologue ($n = 20$)	
Variable	*M*	*SD*	*M*	*SD*	*M*	*SD*
Study 1						
Pretest %	57.0	11.7	55.2	19.8	56.8	17.1
Posttest % (adjusted[a])	81.2 (80.9)	12.0	81.7 (82.6)	13.2	72.8 (72.6)	18.0
Pure gain[b]	24.2	10.9	26.5	13.1	16.0	12.0
	Observing dialogue ($n = 16$)		Tutoring ($n = 8$)		Observing monologue ($n = 16$)	
Study 2						
Pretest %	40.8	14.4	44.6	17.2	40.2	16.9
Posttest % (adjusted[a])	50.2 (50.6)	19.3	64.7 (62.6)	12.7	49.3 (50.0)	14.9
Pure gain[b]	9.5	15.5	20.1	15.9	9.1	11.6

[a] The values in parentheses report the posttest % adjusted by the pretest covariate. [b] Pure gain = unadjusted posttest % – pretest %.

Impact of condition on similar and transfer test questions. Prior work has suggested that students perform better on problems that share a high degree of similarity with the instructional materials, compared to problems that are not similar and so require transfer (Chi, Slotta & de Leeuw, 1994; Gick & Holyoak, 1983; Reed, Dempster, & Ettinger, 1985). Given that transfer is challenging to achieve, an especially powerful intervention would show an effect for transfer test questions. We now analyze whether this was the case in our experiment.

A common way to operationalize transfer is to classify a test item as involving transfer if its cover story is based on a novel problem situation, one that is substantially different from the instructional materials given to the students (Novick & Holyoak, 1991; Reed, 1987; Schwartz & Bransford, 1998). In our experiment, the instructional materials corresponded to the diffusion text and the problem situations covered by the diffusion workbook. Thus, we labeled test questions as *transfer* questions if their cover story included a novel situation, one not mentioned in these materials. For instance, one of the multiple-choice questions asked students how a towel hung on a clothesline gets dry, the correct choice reflecting that the water molecules randomly bounce out of the towel (instead of, for instance, that they are pulled out by the concentration gradient). While this situation was never mentioned in any of the instructional materials, these materials did provide information on random molecular movement, and so a student could answer this and other transfer questions by generating additional inferences necessary to adapt to the novel situation. The remaining "nontransfer" test questions all included a cover story similar to that in the instructional materials, such as the "dye diffusing in water" situation illustrated in the text and workbook, and were labeled as *similar*. (Note that if a tutor or tutee deviated from the instructional materials and, for instance, introduced a situation not in the workbook that was in the test, this could impact the labels. This never occurred, as we verified by checking the content of the videos to ensure our labels were correct.) There were 11 transfer posttest questions (with nine corresponding pretest questions) and 18 similar posttest questions (with 16 corresponding pretest questions).

The data are shown in Table 3 (top), including the means and standard deviations for the pretest, posttest, and gain scores in each condition for the transfer and similar test questions (the transfer and similar pretest scores were calculated using pretest questions whose corresponding counterparts were labeled as transfer and similar in the posttest, respectively).

We stated above that prior research indicated transfer test questions to be more difficult than similar questions for students. This was also the case in our data. As shown in Table 3, the adjusted transfer posttest scores were lower than the similar posttest scores; the transfer pretest scores were also lower compared to the similar pretest scores. Thus, although our criterion for determining whether a question was one of transfer or not depended on whether it was embedded in a context similar to one used in the instructional materials, the lower overall pretest scores across conditions for the transfer questions also indicate that these questions were in fact harder.

Table 2
Summary of Key Results for Studies 1 and 2

Study	F	t	p	df	η^2	d
Study 1	4.23		.02	46	.09	
Planned comparisons						
Dialogue vs. tutoring		0.42	.68	28		0.16
Dialogue vs. monologue		2.46	.02	38		0.78
Study 2	2.80		.07	36	.08	
Planned comparisons						
Dialogue vs. tutoring		2.11	.04	22		0.92
Dialogue vs. monologue		0.12	.91	30		0.04

For the transfer test questions, an ANCOVA with pretest percentage as the covariate and transfer posttest percentage as the dependent variable found a marginal effect of condition, $F(2, 46) = 2.87$, $p = .07$, $\eta^2 = .09$. Planned comparisons showed that there was no significant difference in adjusted posttest performance for these questions between the dialogue and tutoring conditions (73.3% and 75.6%, respectively), $t(28) = 0.29$, $p = .77$, $d = 0.11$. Moreover, students who observed dialogue had a significantly higher adjusted posttest percentage than did students who observed monologue (73.3% and 60.3%, respectively), $t(38) = 2.05$, $p = .046$, $d = 0.65$. For the similar test questions, an ANCOVA with pretest percentage as the covariate and similar posttest percentage as the dependent variable did not find a significant effect of condition, $F(2, 46) = 2.30$, $p = .11$, $\eta^2 = .05$.

Summary. In Study 1, university students learned about a conceptual emergent topic, namely, diffusion. Despite the fact that diffusion instruction begins in middle school, our participants still found this process difficult to understand. The lack of ceiling at pre- or posttest across conditions is consistent with the literature indicating that emergent misconceptions are deeply rooted (Chi, 2005; Meir et al., 2005). We did, however, find that students learned better in some instructional contexts than others and, in particular, that students learned significantly more from observing dialogue than observing monologue. Our analysis also showed no significant difference in learning from collaborative observation of dialogue and tutoring, coupled with a very small effect size for this comparison.

Study 2

In Study 2, we investigated whether the results from Study 1 would generalize to a younger population. Therefore, Study 2 used Study 1 methodology and similar materials but involved middle school students. As far as the materials are concerned, we have used versions of the diffusion tests in prior studies with middle school students (Muldner et al., 2011) and found them to be appropriate. Keeping the materials similar across the two studies controlled for this variable, making it possible to more accurately measure the effect of instructional context.

Materials, Participants, Design, and Procedure

As mentioned above, the materials for Study 2 were based on Study 1 materials but with some minor refinements. For instance, Study 2 took place at the students' school, and since the school imposed a time restriction, the tests used were shortened and modified slightly to 21 pretest questions and six additional questions for the posttest (total 27 posttest questions; 24 of the 27 posttest questions were taken from the Study 1 posttest).

The workbook also covered the same overall concepts but included some differences. Specifically, two of the seven main problems were new (and replaced two of the Study 1 problems), and the remaining five problems included minor differences, such as an extra activity;

Table 3
Means and Standard Deviations for Each Condition on Pretest, Posttest, and Pure Gain for the Similar and Transfer Test Questions in Studies 1 and 2

	Observing dialogue (n = 20)				Tutoring (n = 10)				Observing monologue (n = 20)			
	Similar		Transfer		Similar		Transfer		Similar		Transfer	
Variable	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Study 1												
Pretest %	60.0	9.8	51.1	21.5	60.6	15.6	45.6	32.5	59.4	18.2	51.7	22.0
Posttest % (adjusted[a])	85.8 (85.6)	10.0	73.6 (73.3)	20.8	86.1 (87.0)	11.8	74.6 (75.6)	20.0	80.3 (80.1)	17.2	60.5 (60.3)	26.4
Pure gain[b]	25.8	11.9	22.5	13.6	25.5	11.5	29.0	21.2	20.9	10.7	8.8	20.3
	Observing dialogue (n = 16)				Tutoring (n = 8)				Observing monologue (n = 16)			
Study 2												
Pretest %	49.4	20.2	31.3	17.5	58.1	21.4	30.0	18.2	51.2	23.2	28.1	14.2
Posttest % (adjusted[a])	57.5 (57.9)	22.3	41.2 (41.4)	18.4	74.6 (72.0)	17.5	52.5 (50.9)	11.4	60.0 (60.9)	17.4	35.9 (36.5)	18.7
Pure gain[b]	8.1	24.0	9.9	13.2	16.5	18.2	22.5	16.5	8.9	13.6	7.8	16.9

[a] The values in parentheses report the posttest % adjusted by the pretest covariate. [b] Pure gain = unadjusted posttest % – pretest %.

the overall number of individual subproblem questions across both studies was comparable, at 20 and 18 for Study 1 and Study 2, respectively.

We created a new set of dialogue and monologue videos for Study 2 using the protocol from Study 1. Eight dialogue videos were created, using four of the tutors who participated in Study 1. Each tutor worked individually and on separate occasions with two middle school tutees (one female) to complete the diffusion workbook. The average length of a dialogue video was 33 min. Each tutor also generated two lecture-style monologues before or after each of his or her dialogue sessions, and these became the eight monologue videos used in Study 2. The average length of a monologue video was 24 min.

The student participants were 40 middle school students (four Grade 7 and 36 Grade 8 students; 22 female), recruited from a Title I school located in a major southwestern U.S. city. All participants were performing at least at grade level. None of the participants had been taught about molecular diffusion at school (as reported by the school principal and the Grade 8 science teacher). Students participated on site at their school but outside of their regular classroom activities and were reimbursed $20 for their time.

The design for Study 2 followed the one used for Study 1, namely, a between-subjects design with three conditions (*collaboratively observing dialogue, one-on-one tutoring, collaboratively observing monologue*), as did the random assignment and study procedure (participants read the diffusion text, completed the pretest, completed the diffusion workbook in one of the three conditions, and completed the posttest). As was the case in Study 1, all sessions were conducted individually in a private room, and the observers worked in same-gender pairs. The one deviation was that since we were in a school setting, we needed to have tighter time control, in that all observers had to finish within 2 hr. All but one of the dialogue videos were slightly longer, so we anticipated that students in the dialogue condition might take a little longer. Thus, we used a yoking procedure in which we ran the longer video first (e.g., dialogue) and yoked the observers in the other condition (e.g., monologue) by how long the first-video observers took.[2] The average length of an observing session was 36 min (35 and 38 min for monologue and dialogue, respectively; p = .2).

Results

Table 1 (bottom) shows the means and standard deviations for the pretest, posttest, and gain scores in each condition (N = 40; eight tutees, 16 dialogue observers and 16 monologue observers). An ANOVA with the pretest data did not reveal significant differences among the conditions, $F(2, 37) = 0.21$, $p = .81$. The average pretest score ranged from 40.2% to 44.6% across the three conditions (see Table 1, bottom), which is lower than the Study 1 pretest scores, which were in the 55%–57% range (see Table 1, top).

Impact of condition on learning. An ANCOVA with pretest percentage as the covariate and posttest percentage as the dependent variable showed a marginally significant effect of condition, $F(2, 36) = 2.80$, $p = .07$, $\eta^2 = .08$; key results for Study 2 are summarized in Table 2 (bottom). Planned comparisons indicated that students who observed dialogue had significantly lower adjusted posttest scores than did students who were tutored (50.6% vs. 62.6%, respectively), $t(22) = 2.11$, $p = .04$, $d = 0.92$. There was no significant difference in adjusted posttest performance between the dialogue and monologue observers (50.6 and 50.0, respectively), $t(30) = 0.12$, $p = .91$, $d = 0.04$.

Impact of condition on similar and transfer test questions. To determine if conditional differences existed for various types of test questions, we applied the approach used for Study 1 to label test

[2] For a dialogue/monologue video pair (i.e., generated by a given tutor), we showed the longer video first (typically dialogue). The other video was then used in the corresponding condition, and participants in that condition were yoked by how long observers watching the first video took (within a 5-min grace period to allow for minor variations). If students took longer outside of the grace period, they were cut off, and if they took less time outside of the grace period, they were asked to review their work. For instance, given a dialogue video of length 30 min and a corresponding monologue video of length 25 min, we would show the dialogue video first. If the observers in that condition took 35 min total to watch the video and fill in the workbook, we would then run the monologue condition, informing the students that they had about 35 min to complete the task. It turned out by chance that two pairs from the monologue condition took longer than the allotted time period and had to be cut off before they finished the workbook problems (but both pairs were on the last problem of the workbook).

questions as *transfer* or *similar*. The results are shown in Table 3 (bottom), including the means and standard deviations for the pretest, posttest, and gain scores in each condition for each category. As in Study 1, in Study 2 the transfer questions also were more challenging than the similar questions: Students' pretest and adjusted posttest scores were lower for the transfer questions (see Table 3).

As shown in Table 3 (bottom), for both question categories, the tutees had higher posttest scores than did the dialogue observers, while the observers' scores were comparable between the dialogue and monologue conditions. However, the ANCOVA did not reach significance for either the transfer questions, $F(2, 36) = 2.20$, $p = .13$, $\eta^2 = .08$, or for the similar questions, $F(2, 36) = 2.26$, $p = .12$, $\eta^2 = .07$.

Summary. Study 2 did not replicate the results of Study 1. In Study 2 students learned less from collaboratively observing dialogue than from being tutored, and there was no significant difference between the dialogue and monologue conditions. We now carry out further analysis to shed some light on the discrepancy between the Study 1 and Study 2 results.

Reconciling the Results From Study 1 and Study 2

Some researchers have proposed that a large portion of the tutoring benefit is due to the fact that tutees are interactive with the tutor (Chi et al., 2001, 2008), which provides constructive opportunities for the tutee. For instance, tutor prompting and scaffolding encourages tutees to generate *substantive contributions*, which are domain-related utterances that have been shown to be positively associated with learning (Chi et al., 2008). Collaborative observers also have constructive opportunities, but their interaction is not driven by a tutor. This difference may have various implications in terms of how student behaviors impact learning in general and the production and benefit of substantive contributions in particular. Thus, we now analyze students' substantive contributions during observing and being tutored and speculate on how the findings can be used to interpret and reconcile our results from Study 1 and Study 2.

Generation of Substantive Contributions

To identify student substantive contributions, we followed the convention used in the Chi et al. (2008) work by analyzing students' verbal utterances. Thus, we first segmented the protocols at the phrase level and then relied on the following definition from Chi et al.:

> A substantive segment is defined as a meaningful contribution to an ongoing activity, such as problem solving, or a relevant response to a tutor's explanations. (p. 325)

While Chi et al. used this definition to identify tutees' substantive contributions during a tutoring session, the definition can also be used to label observers' contributions during collaborative observation. Given this definition, segments related to the domain after students read the target problem statement were considered substantive. We did not consider simply reading a problem statement in the diffusion workbook as substantive, for two main reasons. First, we did not ask students to read the problem statements out loud, and some read them silently, which could have biased the coding. More important, for this coding, we wanted our analysis to go beyond a student being on task, by capturing student ideas and thoughts on diffusion that were not exactly based on the instructional materials. Note, however, that students could refer to the problem statement in their own words while talking to their partner, and that would be considered substantive. Other segments that were not considered substantive included simple agreement (e.g., "yeah" or "uh huh"), repetition, or off-task remarks. To illustrate, consider the following single student utterance, which includes four segments:

[1] If there is a flow of dye, which direction would it appear to move and why //

[2] I think it would be in all directions //

[3] because the guy said that you wouldn't be able to know where they go //

[4] because they are always moving.

The first segment corresponds to the student reading a diffusion workbook question and so, according to the coding scheme, is not considered substantive. The last three segments (Segments 2–4) are considered substantive because they describe how the molecular motion is random and continuous.

Two researchers coded a random portion of the observer and tutee transcripts (corresponding to 20% of the transcripts for Study 1 and Study 2) for substantive contributions. Interrater reliability was high (Kappa = .88). Thus, the remainder of the protocols was coded by one of the researchers.

Overall, collapsing across conditions and studies, substantive contributions were positively correlated with posttest scores (after controlling for pretest, $r = .34$, $p < .01$), which replicates results from prior work (Chi et al., 2008). Furthermore, collapsing across conditions, university students generated more substantive contributions than did middle school students (46.40 and 31.87, respectively), $t(84) = 2.3$, $p = .02$, $d = 0.51$.[3] As reported by an ANCOVA with pretest as the covariate, the college students also had higher adjusted posttest scores than did the middle school students (77.93% vs. 52.75, respectively), $F(1, 87) = 26.22$, $p < .01$, $\eta^2 = .11$. We then performed more fine grained analysis exploring conditional effects on substantive contribution generation.

Number of Substantive Contributions

Table 4 shows the means and standard deviations for the number of substantive contributions in the three conditions for Study 1 and Study 2. In both studies, the effect of contributions was significant, as indicated by an ANOVA with substantive contributions as the dependent variable: Study 1: $F(2, 45) = 19.36$, $p < .01$, $\eta^2 = .46$; Study 2: $F(2, 35) = 38.45$, $p < .01$, $\eta^2 = .69$. In Study 1, the tutees generated more substantive contributions than did the dialogue observers (87.1 vs. 45.4, respectively), $t(27) = 4.44$, $p < .01$, $d = 1.78$, and the dialogue observers produced more substantive contributions than did the monologue observers (45.4 vs. 28.2, respectively), $t(37) = 2.29$, $p = .03$, $d = 0.73$. In Study 2, the tutees also generated more substantive contributions than did the dialogue observers (75.0 vs. 23.7, respectively), $t(21) = 7.84$, $p < .01$, $d = 3.55$, but there was no significant difference between the dialogue

[3] Slight variations in *df* are due to the removal of two outliers in the substantive data, which were identified by the SPSS application (one in the monologue condition and one in the tutoring condition for Study 1 and for Study 2).

Table 4
Means and Standard Deviations for Each Condition for Total Substantive Contributions in Study 1 and Study 2

	Observing dialogue			Tutoring			Observing monologue		
Study	*M*	*SD*	*N*	*M*	*SD*	*N*[a]	*M*	*SD*	*N*[a]
Study 1	45.4	26.1	20	87.1	31.8	9	28.2	14.2	19
Study 2	23.7	11.3	16	75.0	25.8	7	20.5	10.0	15

[a] Slight variations in *N* are due to outlier removal.

and monologue observers (23.7 and 20.5, respectively), $t(29) = 0.62$, $p = .54$, $d = 0.22$.

Given that the generation of substantive contributions has been shown to be associated with learning (Chi et al., 2008), this data provides a possible explanation for our observer-related results. That is, for Study 1, the dialogue observers generated more substantive contributions than did the monologue observers, and this pattern corresponds to their learning, in that the dialogue observers learned more than did the monologue observers. Similarly, in Study 2 there was no significant difference in the number of substantive contributions between the dialogue and monologue conditions, which could explain the lack of a reliable difference in student learning between the observing conditions in Study 2.

The generation of substantive contributions may also explain the superior learning of the tutees in Study 2 compared to the dialogue observers, since the pattern of learning and the pattern of substantive contributions correspond. However, this correspondence is not apparent for the Study 1 tutees. That is, for Study 1, even though the tutees generated significantly more substantive contributions than did the dialogue observers, the tutees did not learn more than these observers. The next analysis attempts to explain this dilemma.

Analyses of Tutors' Elicitation Style and Its Impact on Substantive Contributions

We now speculate on the style of interaction between the individuals within a pair (a tutor and tutee vs. two collaborative observers) as a factor in the suppressed advantage of the tutees' substantive contributions compared to the observers' substantive contributions in Study 1. In a tutorial session, the tutor is driving the interaction through scaffolding prompts and questions while the student responds to these prompts and questions, an interaction style we refer to as *question-and-answer* below. To illustrate, here is an excerpt of a typical tutoring session between a tutor *T* and a student *S* (Study 1):

> [1] *T:* Ok, so describe—this is kind of a redundant question but describe how the water and dye molecules are behaving.
>
> [2] *S:* Alright um . . . the molecules they are continually moving in a random . . . aaa . . . direction, most likely straight and when they hit, they bounce off each other and move off in that direction.
>
> [3] *T:* Right so if I were to ask you to draw arrows for this water molecule, how would you draw it?
>
> [4] *S:* Umm . . . just like a general arrow?
>
> [5] *T:* So the same thing we did for this one (points).
>
> [6] *S:* Ok so for example it could move this way or move towards this one and then bounce off . . .
>
> [7] *T:* And what other directions could it go?
>
> [8] *S:* It can go this way and that way.
>
> [9] *T:* Right so when we just compare the arrows then it looks like they are behaving in exactly the same way, right?

The vast majority of tutees' substantive contributions were elicited by their tutor in this manner, highlighting that tutees rarely initiate information without being prompted by a tutor question (on average, only 13.6% of the tutee substantive contributions were initiated by the tutees in Study 1 and 3.2% in Study 2).

Given that the tutoring condition had the question-and-answer style of interaction, it is not surprising that the tutees generated many substantive contributions, since the tutees' contributions were mostly elicited by a tutor. However, in Study 1, the tutee substantive contribution frequency did not result in the tutees' learning significantly more than did the dialogue observers. Why would more not be better, as far as tutees' substantive contributions are concerned? This result supports the "interaction plateau" conjecture (VanLehn, 2011), which suggests that the benefits of interaction between a tutor and a tutee eventually level off, reaching a point where more interactivity does not improve learning. Since a byproduct of interactivity is the generation of substantive contributions, it may be that the Study 1 tutee participants encountered this plateau because they had generated so many substantive contributions (more than did the observers), and subsequent generation of substantive contributions did not produce more learning. But why is a plateau encountered?

One possibility relates to the fact that the contributions produced by the tutees are not always adapted to their individual needs, since the tutor elicits such contributions. In some sense, the tutees rely on the tutor to elicit the information that helps them learn. However, tutors are not very adept at accurately assessing the mental models of their tutees (Chi, Siler, & Jeong, 2004; Herppich, Wittwer, Nückles, & Renkl, 2013), suggesting that at least some of the time the tutor may be asking questions that do not foster learning. For instance, tutors may be eliciting information that is already known by their tutees, which would not be as beneficial for learning and could explain the interaction plateau. To see if this was the case, we analyzed tutees' substantive contributions during their tutoring sessions.

We labeled a tutee's substantive contribution as *incorrect* if it corresponded to an incorrect tutee contribution during the tutorial session or contribution(s) directly following an incorrect response and related to it (i.e., corresponding to the tutor's probing the student about his or her incorrect response). Remaining substantive contributions were labeled as *correct*. For Study 1, we found that, on average, 74.5% of the tutee contributions were *correct*. One interpretation of this result is that the Study 1 tutees may have already known much of what the tutor was eliciting, thereby providing an explanation for why they did not learn more than did the dialogue observers even though they generated many more substantive contributions. This result might also explain the interaction plateau.

In contrast to the Study 1 tutees, the Study 2 tutees did learn significantly better than did the dialogue observers. Why was this the case? To answer this question, we examined whether the elicited tutee contributions for Study 2 were mostly *incorrect*, thus accounting for the greater learning of the tutees compared to the dialogue observers in that study. When we analyzed the *correct* versus *incorrect* substantive contributions for the Study 2 tutees, we found that 39.5% were *correct*, a much lower proportion than in Study 1 (74.5%). The difference between Study 1 and Study 2 *correct* proportions is significant, $t(16) = 5.86$, $p < .01$, $d = 2.78$, and suggests that in contrast to the Study 1 tutees, the Study 2 tutees may not have reached the interaction plateau, because many of the contributions they provided corresponded to concepts they still had to learn. This in turn provides a clue as to why, in Study 2, the tutees learned significantly more than did the dialogue observers. A complementary explanation relates to the fact that the Study 2 dialogue observers did not generate many substantive contributions, significantly less than did the Study 1 dialogue observers (23.7 and 45.4, respectively), $t(34) = 3.3$, $p < .01$, $d = 1.5$.

Exploratory regression analyses. The above analyses and interpretations suggest that substantive contributions may have a different impact for observers than for tutees, since so many of the tutee contributions are elicited instead of self-generated. To substantiate this interpretation, we analyzed the relationship of substantive contributions and posttest performance for tutoring versus observing more closely, by running an exploratory linear regression, with posttest as the dependent variable and pretest, substantive, condition and Condition × Substantive interaction as the explanatory variables. Because we were primarily interested in analyzing how being substantive influenced learning in the tutoring versus observing contexts, without distinguishing dialogue and monologue, we collapsed the two observing contexts, so that condition had two levels: observing and tutoring. Since condition is a categorical variable, we transformed it into two binary dummy variables (0, 1), where 1 corresponded to observing and 0 corresponded to tutoring. Note that with a "dummy" coding, a given dummy variable represents a comparison between the target variable and the reference variable and so, given N dummy variables, only $N - 1$ can be input into the linear regression. We chose tutoring as the reference variable (because we had only two variables, this choice was arbitrary, given that our focus was on the interaction term, as we describe below).

For Study 1, the overall model we obtained, shown in Table 5 (top), is significant ($R^2 = .63$), $F(4, 43) = 18.07$, $p < .01$. Of primary interest is the interaction term, which reveals whether a difference exists in how substantive contributions influenced posttest scores between the tutoring and observing conditions. Since the interpretation of the other coefficients is affected by the interaction term (Braumoeller, 2004), which essentially renders them "baseline" slopes (Grace-Martin, 2000), they are not relevant and are not discussed here. The interaction is modest but significant and indicates that overall, *substantive* has a stronger positive relationship to posttest for observing compared to tutoring. For Study 2, the overall model we obtained, shown in Table 5 (bottom), is also significant ($R^2 = .49$), $F(4, 33) = 8.0$, $p < .01$. As was the case for Study 1, the interaction term is significant and indicates that substantive has a stronger positive relationship to posttest for observing compared to tutoring.

General Discussion

One of our research goals was to compare the pedagogical utility of being tutored against collaborative observation of dialogue. This is an especially important comparison, given that tutoring is very effective at fostering learning but is impractical in terms of scalability, since providing a human tutor for every student is not practical. Thus, finding alternative interventions that are beneficial for learning but also scalable has been a long-standing goal in the educational psychology field. Our findings indicate that for a university population, collaborative observation of dialogue has higher overall utility than does being tutored, for reasons we expand on now.

When evaluating the pedagogical utility of an intervention, it is paramount to consider not only its *scalability,* in terms of how easily it may be implemented, and its *impact* on learning, but also the *effect size* of that impact (Breaugh, 2003; Hattie, 1999). For instance, a treatment that has a large positive effect may have a high utility even if it is not scalable, while a treatment that is scalable may be desirable even when it is not superior in terms of impact and effect size over a less scalable intervention. What is the threshold for a "meaningful" effect size? While P. A. Cohen et al. (1982) indicated a generic threshold of $d = 0.2$ as a small effect, for educational applications Hattie (1999) argued that in order for an intervention to have practical meaning the effect size needs to be $d = 0.4$ or greater.

For Study 1, the sample effect size of the *tutoring – observing dialogue* comparison was very small at $d = 0.16$, substantially below the Hattie (1999) practical level of $d = 0.4$. Thus, even though the lack of a significant difference for this comparison means we cannot determine which treatment was superior, we can still conclude that

Table 5
Linear Regression Coefficients for Study 1 and Study 2

Predictors	Unstandardized coefficient	Standardized coefficient	*t*	*p*
Study 1				
Condition × Substantive	0.30	.49	2.33	.02
Constant	46.03	—	4.18	.00
Condition	−17.40	−.45	−1.65	.11
Pretest	0.69	.69	7.35	.00
Substantive	−0.05	−.10	−0.49	.66
Study 2				
Condition × Substantive	0.63	.49	2.10	.04
Constant	61.42	—	3.88	.00
Condition	−45.12	−1.06	−2.63	.01
Pretest	0.67	.62	4.77	.00
Substantive	−0.36	−.54	−1.72	.09

observing dialogue had higher utility than did being tutored in Study 1. This is because one-on-one human tutoring is a much less scalable intervention than observing dialogue, and the sample effect size for this comparison was too low to be meaningful.

Dialogue observation also had a higher utility than did monologue observation in Study 1. This is because although both are scalable approaches, we found that observing dialogue fostered significantly more learning, with a large effect size ($d = 0.78$). Thus, we confirmed results from prior work with scripted content that dialogue observation is superior, as well as generalized this result to an emergent domain. To explain why dialogue was superior, we analyzed students' substantive contributions from the observation sessions and showed that the dialogue observers generated more such contributions than did the monologue observers. In a sense, one could argue that the dialogue observers were more engaged, something that Craig et al. (2009) also found when they compared dialogue and monologue observation. This increased engagement may occur because observers mirror a highly interactive tutoring session, a conjecture proposed by Chi (2013).

The benefits of dialogue observation in Study 1 did not transfer over to Study 2, which involved a younger population. In particular, the Study 2 tutees learned more than did the dialogue observers, with a large effect size ($d = 0.92$); thus, the utility of being tutored was higher in this study. One explanation for this result we proposed above is that the Study 2 dialogue observers did not generate many substantive contributions. A second complementary explanation pertains to student knowledge. Although participants in both our studies found the target domain challenging, as indicated by a lack of ceiling at pretest or posttest, the domain may have been particularly difficult for the younger Study 2 population, as suggested by our analysis of the tutee known versus unknown contributions (as well as the lower test scores). Interacting with a tutor is especially beneficial for low-knowledge students (VanLehn et al., 2007), which provides an additional explanation for why the Study 2 observers learned less than did the Study 2 tutees.

Also, in contrast to Study 1 findings, in Study 2 the utility of dialogue and monologue observation was comparable, in that the dialogue observers did not learn more than did the monologue observers, and the sample effect size was very small for this comparison ($d = 0.04$). Our analysis of substantive contributions again provides an explanation for this finding, since there was no significant difference between the two observing conditions in terms of these contributions. Thus, the benefit of dialogue encouraging observers to be more substantive that we saw in Study 1 did not transfer to Study 2. Why would this be the case?

It is possible that the younger students may have lacked the necessary metacognitive skills, for instance needed to realize lack of understanding, which could be repaired through the generation of substantive contributions during discussion with one's partner. There is some evidence that metacognitive skills increase as students progress from grade school to university (Veenman, Wilhelm, Beishuizen, 2004). Thus, younger students may require additional scaffolding to contribute substantively in a collaborative situation, over and beyond the implicit support offered by a dialogue video.

An additional explanation for the difference in Study 1 and Study 2 observer results could be related to the content of the Study 1 and Study 2 videos, if only the Study 1 dialogue videos contained more of those beneficial features (deep questions, misconceptions) that we indicated in the introduction aid observer learning. This, however, was not the case: Both the Study 1 and Study 2 dialogue videos contained more deep questions than did the monologue videos (Study 1: 18.2 vs. 2.3, respectively; Study 2: 25.2 vs. 1.8, respectively) and more refuted misconceptions than did the monologue videos (in a monologue video, a refuted misconception corresponded to a tutor stating an incorrect concept and then refuting it; Study 1: 5.9 vs. 2.5, respectively; Study 2: 9.0 vs. 2.4, respectively). Thus, we cannot draw conclusions as to the impact of these features on observer learning in our studies, since only the Study 1 dialogue observers followed the expected pattern of learning more in the presence of these features. However, in contrast to prior work exploring the role of questions or misconceptions (e.g., Craig et al., 2004; Driscoll et al., 2003; Muller et al., 2007; Schunk & Hanson, 1985), in our experiments students worked collaboratively. Such a context may diminish some of the misconception and question effects, because the effect of collaboration may be stronger and so may overshadow other effects. Moreover, a student's collaborator may generate behaviors associated with these features. For instance, observers don't have ideal domain knowledge and consequently may produce misconceptions or generate questions to their partner. Thus, more work is needed to understand how these features (questions, misconceptions) influence learning in collaborative situations.

Limitations

In our studies, the observing conditions had more participants than did the tutoring condition, because we wanted (1) to balance the content by having each tutee video be viewed by exactly one pair of dialogue observers and (2) to afford the observers interaction opportunities by having them work in pairs. While Howell (2010) proposed that unequal sample sizes across conditions can easily be accounted for in a one-way design (e.g., supported by SPSS and used in our analysis), such a design could result in a loss of power, and so our results should be replicated with additional studies and larger subject pools.

A second limitation pertains to the post hoc nature of our analysis used to explain our results from Study 1 and Study 2, related to the generation of substantive contributions in a collaborative setting. Such contributions are a form of active and/or constructive behaviors, which prior work has shown to be highly beneficial, as conceptualized in the interactive-constructive-active-passive framework (Chi, 2009). The majority of our results mirror this pattern, in that conditions in which students produced more contributions also had more learning, but we did not experimentally control for this factor a priori. Consequently, our results, while consistent with our findings related to student learning, need to be replicated with controlled experiments where substantive contributions are explicitly manipulated.

Future Directions and Implications

Since Study 2 showed middle school students did not learn optimally from collaboratively observing dialogue, future work should explore how to better support younger students in order to maximize their learning. One approach is to devise scaffolding that will help students generate more substantive contributions, such as prompts embedded in the video reminding students to reflect (Ogan, Aleven, & Jones, 2009; van Blankenstein, Dolmans, van der Vleuten & Schmidt, 2011). In addition, we plan to explore how students' prior knowledge impacts learning from collaborative observation, which may shed further light on why, compared to Study 1, Study 2 observers' learning was lower.

To conclude, we have shown that collaboratively observing dialogue is an effective instructional intervention for difficult conceptual topics in an emergent domain for a university-level population. This paradigm had higher utility than did being tutored, since it fostered more learning and is more scalable. We believe that learning from observing dialogue has the potential to also benefit younger students, but more work is needed to understand how to best support or guide this population during observation. Of course we are not proposing that collaborative observation of dialogue should replace being tutored in all instances, but rather that it can serve as a viable alternative when a tutor is not available, as is the case in many situations. For instance, one natural context for observational learning is online applications, where dissemination is facilitated by the recent surge in E-learning, which has given rise to various courses, degree programs, and learning academies. Currently, online materials are often monologue based (e.g., instructional videos with a "talking head"; Caspi, Gorsky, & Privman, 2005; Zhang, Zhou, Briggs, & Nunamaker, 2006). Thus, these venues are missing an opportunity to optimize student learning with dialogue-based materials for collaborative observers.

References

Aleven, V., McLaren, B., Roll, I., & Koedinger, K. (2006). Toward metacognitive tutoring: A model of help seeking with a cognitive tutor. *International Journal of Artificial Intelligence in Education, 16,* 101–128.

Anderson, J. R. (1993). *Rules of the mind.* Hillsdale, NJ: Erlbaum.

Arroyo, I., Beal, C. R., Murray, T., Walles, R., & Woolf, B. P. (2004). Web-based intelligent multimedia tutoring for high stakes achievement tests. In J. C. Lester, R. M. Vicari, & F. Paraguaçu (Eds.), *Proceedings of the 7th International Conference on Intelligent Tutoring Systems* (pp. 468–477). New York: Springer.

Bandura, A., Ross, D., & Ross, S. A. (1961). Transmission of aggression through imitation of aggressive models. *Journal of Abnormal and Social Psychology, 63,* 575–582. doi:10.1037/h0045925

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13,* 4–16. doi:10.3102/0013189X013006004

Brace, N., Kemp, R., & Snelgar, R. (2003). *SPSS for psychologists: A guide to data analysis using SPSS for Windows* (2nd ed.). Mahwah, NJ: Erlbaum.

Braumoeller, B. F. (2004). Hypothesis testing and multiplicative interaction terms. *International Organization, 58,* 807–820. doi:10.1017/S0020818304040251

Breaugh, J. A. (2003). Effect size estimation: Factors to consider and mistakes to avoid. *Journal of Management, 29,* 79–97. doi:10.1177/014920630302900106

Caspi, A., Gorsky, P., & Privman, M. (2005). Viewing comprehension: Students' learning preferences and strategies when studying from video. *Instructional Science, 33,* 31–47. doi:10.1007/s11251-004-2576-x

Chi, M. T. H. (2005). Common sense conceptions of emergent processes: Why some misconceptions are robust. *Journal of the Learning Sciences, 14,* 161–199. doi:10.1207/s15327809jls1402_1

Chi, M. T. H. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science, 1,* 73–105. doi:10.1111/j.1756-8765.2008.01005.x

Chi, M. T. H. (2013). Learning from observing an expert's demonstration, explanations and dialogues. In J. J. Staszewski (Ed.), *Expertise and skill acquisition: The impact of William G.* Chase (pp. 1–28). New York, NY: Psychology Press.

Chi, M. T. H., Bassok, M., Lewis, M., Reimann, P., & Glasser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13,* 145–182. doi:10.1207/s15516709cog1302_1

Chi, M. T. H., Kristensen, A. K., & Roscoe, R. (2012). Misunderstanding emergent causal mechanism in natural selection. In K. Rosengren, S. Brem, & G. Sinatra (Eds.), *Evolution challenges: Integrating research and practice in teaching and learning about evolution* (pp. 145–173). New York, NY: Oxford University Press. doi:10.1093/acprof:oso/9780199730421.003.0007

Chi, M. T. H., Roscoe, R., Slotta, J., Roy, M., & Chase, M. (2012). Misconceived causal explanations for emergent processes. *Cognitive Science, 36,* 1–61. doi:10.1111/j.1551-6709.2011.01207.x

Chi, M. T. H., Roy, M., & Hausmann, R. G. M. (2008). Observing tutoring collaboratively: Insights about tutoring effectiveness from vicarious learning. *Cognitive Science, 32,* 301–341. doi:10.1080/03640210701863396

Chi, M. T. H., Siler, S. A., & Jeong, H. (2004). Can tutors monitor students' understanding accurately? *Cognition and Instruction, 22,* 363–387. doi:10.1207/s1532690xci2203_4

Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science, 25,* 471–533. doi:10.1207/s15516709cog2504_1

Chi, M. T. H., Slotta, J. D., & de Leeuw, N. (1994). From things to processes: A theory of conceptual change for learning science concepts. *Learning and Instruction, 4,* 27–43. doi:10.1016/0959-4752(94)90017-5

Cohen, E. G. (1994). Restructuring the classroom: Conditions for productive small groups. *Review of Educational Research, 64,* 1–35. doi:10.3102/00346543064001001

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, P. A., Kulik, J. A., & Kulik, C. L. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal, 19,* 237–248. doi:10.3102/00028312019002237

Cox, R., McKendree, J., Tobin, R., Lee, J., & Mayes, T. (1999). Vicarious learning from dialogue and discourse: A controlled comparison. *Instructional Science, 27,* 431–458. doi:10.1007/BF00891973

Craig, S., Chi, M. T. H., & VanLehn, K. (2009). Improving classroom learning by collaboratively observing human tutoring videos while problem solving. *Journal of Educational Psychology, 101,* 779–789. doi:10.1037/a0016601

Craig, S., Driscoll, D., & Gholson, B. (2004). Constructing knowledge from dialog in an intelligent tutoring system: Interactive learning, vicarious learning, and pedagogical agents. *Journal of Educational Multimedia and Hypermedia, 13,* 163–183.

Craig, S., Gholson, B., Ventura, M., & Graesser, A. (2000). Overhearing dialogues and monologues in virtual tutoring sessions: Effects on questioning and vicarious learning. *International Journal of Artificial Intelligence in Education, 11,* 242–253.

Craig, S., Sullins, J., Witherspoon, A., & Gloslon, B. (2006). The deep-level-reasoning-question effect: The role of dialogue and deep-level-reasoning questions during vicarious learning. *Cognition and Instruction, 24,* 565–591. doi:10.1207/s1532690xci2404_4

Dimitrov, D. M., & Rumrill, P. D. (2003). Pretest-posttest designs and measurement of change. *Work, 20,* 159–165.

Driscoll, D. M., Craig, S., Gholson, B., Hu, X., & Graesser, A. (2003). Vicarious learning: Effects of overhearing dialog and monologue-like discourse in a virtual tutoring session. *Journal of Educational Computing Research, 29,* 431–450. doi:10.2190/Q8CM-FH7L-6HJU-DT9W

Faul, F. (2012). G*Power (Version 3.1.3). [Computer software]. Kiel, Germany: University of Kiel.

Fox Tree, J. E. (1999). Listening in on monologues and dialogues. *Discourse Processes, 27,* 35–53. doi:10.1080/01638539909545049

Fox Tree, J. E., & Mayer, S. A. (2008). Overhearing single and multiple perspectives. *Discourse Processes, 45,* 160–179. doi:10.1080/01638530701792867

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15,* 1–38. doi:10.1016/0010-0285(83)90002-6

Grace-Martin, K. (2000). *Interpreting interactions in regression*. Retrieved July 1, 2012, from http://www.cscu.cornell.edu/news/statnews/stnews40.pdf

Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology, 9,* 495–522. doi:10.1002/acp.2350090604

Harskamp, E., Ding, N., & Suhre, C. (2008). Group composition and its effect on female and male problem solving in science education. *Educational Research, 50,* 307–318. doi:10.1080/00131880802499688

Hattie, J. (1999). *Influences on student learning*. Retrieved July 1, 2012, from http://coburgmaths.edublogs.org/files/2008/05/1-john-hattie-14th-module-5-reading-1-hattie.pdf

Herppich, S., Wittwer, J., Nückles, M., & Renkl, A. (2013). Does it make a difference? Investigating the assessment accuracy of teacher tutors and student tutors. *Journal of Experimental Education, 81,* 242–260. doi:10.1080/00220973.2012.699900

Howell, D. C. (2010). *Statistical methods for psychology* (7th ed., p.). Belmont, CA: Wadsworth.

Johnson, D. W., & Johnson, R. T. (2009). An educational psychology success story: Social interdependence theory and cooperative learning. *Educational Researcher, 38,* 365–379. doi:10.3102/0013189X09339057

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1995). Intelligent tutoring goes to school in the big city. In J. Greer (Ed.), *Proceedings of the 7th World Conference on Artificial Intelligence and Education* (pp. 421–428). Charlottesville, NC: AACE.

Lepper, M. R., Woolverton, M., Mumme, D. L., & Gurtner, J. L. (1993). Motivational techniques of expert human tutors: Lessons for the design of computer-based tutors. In S. P. Lajoie & S. J. Derry (Eds.), *Computers as cognitive tools* (pp. 75–105). Hillsdale, NJ: Erlbaum.

Levy, S., & Wilensky, U. (2008). Inventing a mid level to make ends meet: Reasoning between the levels of complexity. *Cognition and Instruction, 26,* 1–47. doi:10.1080/07370000701798479

Meir, E., Perry, J., Stal, D., & Klopfer, E. (2005). How effective are simulated molecular-level experiments for teaching diffusion and osmosis? *Cell Biology Education, 4,* 235–248. doi:10.1187/cbe.04-09-0049

Merrill, D. C., Reiser, B. J., Merrill, S. K., & Landes, S. (1995). Tutoring: Guided learning by doing. *Cognition and Instruction, 13,* 315–372. doi:10.1207/s1532690xci1303_1

Muldner, K., & Conati, C. (2010). Scaffolding meta-cognitive skills for effective analogical problem solving via tailored example selection. *International Journal of Artificial Intelligence in Education, 20*(2), 99–136.

Muldner, K., Dybvig, K., Lam, R., & Chi, M. T. H. (2011). Learning by observing tutorial dialogue versus monologue collaboratively or alone. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1340–1346). Austin, TX: Cognitive Science Society.

Muller, D. A., Bewes, J., Sharma, M. D., & Reimann, P. (2008). Saying the wrong thing: Improving learning with multimedia by including misconceptions. *Journal of Computer Assisted Learning, 24,* 144–155. doi:10.1111/j.1365-2729.2007.00248.x

Muller, D. A., Sharma, M. D., Eklund, J., & Reimann, P. (2007). Conceptual change through vicarious learning in an authentic physics setting. *Instructional Science, 35,* 519–533. doi:10.1007/s11251-007-9017-6

Murray, T. (1999). Authoring intelligent tutoring systems: An analysis of the state of the art. *International Journal of Artificial Intelligence in Education, 10,* 98–129.

Novick, L. R., & Holyoak, K. J. (1991). Mathematical problem solving by analogy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 398–415. doi:10.1037/0278-7393.17.3.398

Ogan, A., Aleven, V., & Jones, C. (2009). Advancing development of intercultural competence through supporting predictions in narrative video. *International Journal of Artificial Intelligence in Education, 19,* 267–288.

Reed, S. K. (1987). A structure-mapping model for word problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 124–139. doi:10.1037/0278-7393.13.1.124

Reed, S. K., Dempster, A., & Ettinger, M. (1985). Usefulness of analogous solutions for solving algebra word problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 106–125. doi:10.1037/0278-7393.11.1.106

Resnick, M. (1996). Beyond the centralized mindset. *Journal of the Learning Sciences, 5,* 1–22. doi:10.1207/s15327809jls0501_1

Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and observers. *Cognitive Psychology, 21,* 211–232. doi:10.1016/0010-0285(89)90008-X

Schunk, D. H., & Hanson, R. A. (1985). Peer models: Influence on children's self-efficacy and achievement. *Journal of Educational Psychology, 77,* 313–322. doi:10.1037/0022-0663.77.3.313

Schunk, D. H., Hanson, R. A., & Cox, P. D. (1987). Peer-model attributes and children's achievement. *Journal of Educational Psychology, 79,* 54–61. doi:10.1037/0022-0663.79.1.54

Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. *Cognition and Instruction, 16,* 475–522. doi:10.1207/s1532690xci1604_4

Smith, C., Carey, S., & Wiser, M. (1985). On differentiation: A case study of the development of the concept of size, weight, and density. *Cognition, 21,* 177–237. doi:10.1016/0010-0277(85)90025-3

Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology, 51,* 473–481. doi:10.1037/0022-0167.51.4.473

van Blankenstein, F., Dolmans, D., van der Vleuten, C., & Schmidt, H. (2011). Which cognitive processes support learning during small-group discussion? The role of providing explanations and listening to others. *Instructional Science, 39,* 189–204. doi:10.1007/s11251-009-9124-7

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems and other tutoring systems. *Educational Psychologist, 46,* 197–221. doi:10.1080/00461520.2011.611369

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rosé, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science, 31,* 3–62. doi:10.1080/03640210709336984

VanLehn, K., Lynch, C., Schulze, K., Shapiro, J. A., Shelby, R., Taylor, L., . . . Wintersgill, M. (2005). The Andes Physics Tutoring System: Lessons learned. *International Journal of Artificial Intelligence in Education, 15,* 147–204.

Veenman, M. V. J., Wilhelm, P., & Beishuizen, J. (2004). The relation between intellectual and metacognitive skills from a developmental perspective. *Learning and Instruction, 14,* 89–109. doi:10.1016/j.learninstruc.2003.10.004

Walker, E., Rummel, N., & Koedinger, K. R. (2011). Designing automated adaptive support to improve student helping behaviors in a peer tutoring activity. *International Journal of Computer-Supported Collaborative Learning, 6,* 279–306. doi:10.1007/s11412-011-9111-2

Wang, Y., & Heffernan, N. T. (2011). The "assistance" model: Leveraging how many hints and attempts a student needs. In R. C. Murray & P. M. McCarthy (Eds.), *Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference* (pp. 549–554). Menlo Park, CA: AAAI.

Wuensch, K. (2012). *Power analysis for ANCOVA*. Retrieved from http://core.ecu.edu/psyc/wuenschk/MV/LSANOVA/Power-ANCOV.doc

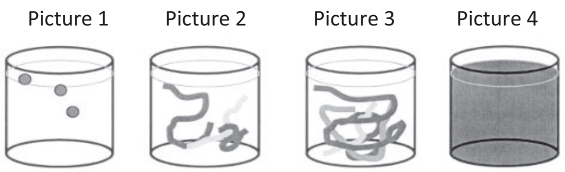
Zhang, D., Zhou, L., Briggs, R. O., & Nunamaker, J. F. (2006). Instructional video in e-learning: Assessing the impact of interactive video on learning effectiveness. *Information & Management, 43,* 15–27. doi:10.1016/j.im.2005.01.004

(*Appendix follows*)

Appendix

Sample Test Questions Used in Study 1 and Study 2

We have a glass full of clear water. We add several drops of dark green dye (Picture 1). You will see that the dye seems to swirl and spread through the water (Pictures 2 and 3). Eventually you see the water–dye solution is a uniform color (Picture 4).



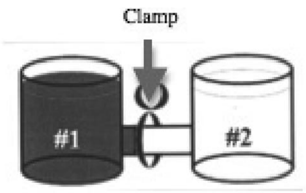
1. [similar] When the dye is dropped into the water, it will appear to flow in the beaker because:

a. the dye molecules are colliding and spreading, while the water molecules stay in place.
b. the dye molecules are colliding and spreading to where there is more room in the water areas of the solution.
c. the dye molecules are dissolving into the water molecules creating the flow pattern.
d. all the molecules are colliding and spreading, leading to changes in concentration from one area to another.

2. [similar] Eventually, the dye doesn't appear to spread anymore and the solution is a uniform color (Picture 4). This is called equilibrium. When equilibrium is reached:

a. all of the molecules stop moving.
b. all of the molecules keep moving.
c. the dye molecules stop moving.
d. the water molecules stop moving.

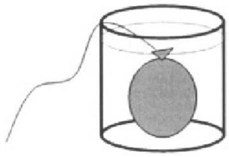
Suppose you have two beakers connected by a short tube with a clamp (see below). Beaker #1 contains a highly concentrated solution of darkly colored blue dye and water. Beaker #2 contains only water.



3. [similar] The clamp between the beakers is removed. Choose the most correct statement:

a. Once equilibrium is reached, we will not see a blue flow of dye because the dye molecules have stopped moving.
b. At equilibrium, the concentrations will be about the same in both beakers so we will no longer see a flow of blue dye.
c. We will start to see a blue flow back into Beaker #1 after equilibrium since the molecules are always moving.
d. Sometimes the dye molecules will attract each other and the flow of dye will start to move back into Beaker #1.

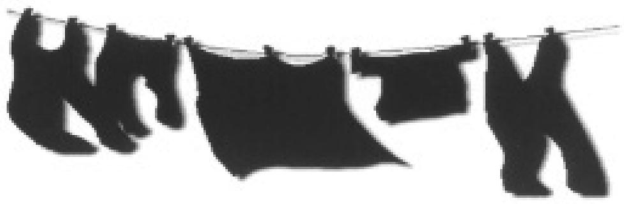
A glass has 100 ml of water mixed with 10 spoons of sugar. A balloon is filled with 50 ml of water and 1 spoon of sugar. Both water and sugar molecules can pass through the balloon walls.



4. [transfer] After you put the balloon in the glass, **eventually equilibrium is reached. At equilibrium:**

a. the sugar molecules in the balloon and the glass exchange places so the sweetness in the glass never changes.
b. sugar molecules will have moved from the balloon to the glass, so the water in the glass will taste sweeter than before.
c. the sugar and water molecules will have spread around, so the water in the glass will taste less sweet than before.
d. the sugar and water molecules will have spread around, so the water in the glass will taste sweeter than before.

When you hang a wet shirt on a clothesline on a sunny day, it dries through the process of evaporation. Evaporation is similar to the process of diffusion.



(*Appendix continues*)

5. [transfer] Which sentence explains how the shirt gets dry?

a. The water molecules in the shirt collide continuously and by chance some bounce from the shirt into the air.
b. The water molecules leave the shirt when the sun's heat breaks the bonds between the molecules and the shirt.
c. The air has a low concentration of water, which pulls the water molecules out of the shirt.
d. Heat from the sun forces the water molecules to expand and this causes the water molecules to leave the shirt.

6. [transfer] If a water molecule leaves the shirt, can it ever go back into the shirt?

a. Yes
b. Yes, if it is attached to another water molecule
c. No
d. No, unless it is attached to another air molecule

Received April 14, 2012
Revision received June 21, 2013
Accepted July 29, 2013 ■