

An Experimental Study of the Effects of Representational Guidance on Collaborative Learning Processes

Daniel D. Suthers and Christopher D. Hundhausen
Laboratory for Interactive Learning Technologies
University of Hawai'i at Manoa

The importance of both social processes and of representational aids for learning is well-established, yet few experimental studies have addressed the combination of these factors. The research reported in this article evaluates the influence of tools for constructing representations of evidential models on collaborative learning processes and outcomes. Pairs of participants worked with 1 of 3 representations (Graph, Matrix, Text) while investigating complex science and public health problems. Dependent measures included (a) the content of participants' utterances and representational actions and the timing of these utterances and actions with respect to the availability of information; (b) a multiple choice test of the ability to recall the data, hypotheses, and evidential relations explored; and (c) the contents of a written essay. The results show that representational notations can have significant effects on learners' interactions, and may differ in their influence on subsequent collaborative use of the knowledge being manipulated. For example, Graph and Matrix users elaborated on previously represented information more than Text users. Representation and discussion of evidential relations was quantitatively greatest for Matrix users as predicted, yet this came at the cost of excessive consideration and revision of unimportant relations. Graph users may have been more focused in their consideration of evidence, and the work done in the Graph representation had the greatest impact on the contents of the essays. Although limited to initial use of representations in a laboratory setting, the work demonstrates that representational guidance of collaborative learning is worthy of study and suggests several lines of further investigation.

The importance of social processes to learning, including the potential utility of collaborative learning, is well established (Brown & Campione, 1994; Lave & Wenger, 1991; Scardamalia & Bereiter, 1991; Slavin, 1990; Webb & Palincsar, 1996). Likewise, prior work has shown the importance of representational aids to individual understanding and problem solving (Koedinger, 1991; Kotovsky & Simon, 1990; Larkin & Simon, 1987; Novak, 1990; Novick & Hmelo, 1994; Zhang, 1997). Yet few studies have addressed the combination of these factors; namely, the role of representational aids in supporting group learning processes. Exceptions include Baker and Lund (1997), Guzdial (1997), and Schwarz, Neuman, Gil, and Ilya (2003). The research reported in this article addresses the question of how we might design representational systems that guide and support collaborative learning processes (such as discourse) in a positive way.

Our study has several converging motivations. Prior work by Suthers and colleagues on Belvedere, a networked environment for collaborative construction of “evidence maps” (Suthers, Toth, & Weiner, 1997; Suthers & Weiner, 1995), suggested that the representational bias of tools such as Belvedere might influence students’ discussion. Concurrently, other researchers were using different representations for similar objectives (supporting epistemic reasoning in science). For example, SenseMaker (Bell, 1997) used a container representation (in which data are sorted into theory containers); WebCamile (Guzdial et al., 1997) and SpeakEasy (Hoadley, Hsi, & Berman, 1995) used threaded discussions; and Puntambekar, Nagel, Hübscher, Guzdial, and Kolodner (1997) studied matrix representations. However, the choice of representations for these systems was generally not based on systematic comparisons of the effects of representations on collaborative learning. Theoretical inspirations for undertaking such a comparison came from Roschelle’s (1994) observation that shared representations (animations and simulations in his case) serve to mediate collaborative inquiry; and from Collins and Ferguson’s (1993) discussion of representations as “epistemic forms” with associated “epistemic games.” Other theoretical motivations are discussed in Suthers (1999b, 2001) as well as in this article. Based on these motivations, Suthers and colleagues undertook a classroom study of representational effects (Toth, Suthers, & Lesgold, in press) that will be discussed in the conclusions of this article. Pragmatic constraints limited that study to analysis of students’ artifacts. We were unable to directly observe the roles representations played in students’ collaborative learning processes, further motivating the laboratory study reported here.

ROLES OF EXTERNAL REPRESENTATIONS IN COLLABORATIVE LEARNING

External representations have long been a subject of study in the context of learning and problem-solving tasks, with research showing that the choice of represen-

tation can influence an individual's conception of a problem and hence the ease of finding a solution (see previous citations). One might ask whether it is sufficient to extrapolate from this work, predicting representational effects on groups by aggregating effects on individuals. Although we believe that much can be gained from such reasoning, we also believe that the shared use of representations by distributed cognitions (Salomon, 1993) might involve additional emergent phenomena. External representations play at least three roles that are unique to situations in which a group is constructing and manipulating shared representations as part of a constructive activity:

1. Initiating negotiations of meaning.
2. Serving as a representational proxy for purposes of gestural deixis.
3. Providing a foundation for implicitly shared awareness.

These roles are discussed here.

1. An individual who wishes to add to or modify a shared representation may feel some obligation to obtain agreement from one's group members, leading to negotiations about and justifications of representational acts. This discourse will include negotiations of meaning and shared belief that would not be necessary in the individual case, where one can simply change the representation as one wishes. The creative acts afforded by a given representational notation may therefore affect which negotiations of meaning and belief take place.

2. The components of a collaboratively constructed representation, having arisen from negotiations of the type just discussed, evoke in the minds of the participants rich meanings beyond that which external observers might be able to discern by inspection of the representations alone. These components can serve as an easy way to refer to ideas previously developed, this reference being accomplished by gestural deixis (reference to an entity relative to the context of discourse by pointing) rather than verbal descriptions (Clark & Brennan, 1991). In this manner, collaboratively constructed external representations facilitate subsequent negotiations, increasing the conceptual complexity that can be handled in group interactions and facilitating elaboration on previously represented information.

3. The shared representation also serves as a group memory, reminding the participants of previous ideas (encouraging elaboration on them) and possibly serving as an agenda for further work. Individual work also benefits from an external memory, but in the group case there is an additional awareness that one's interlocutors may be reminded by the representation of prior ideas, prompting oneself to consider potential commentary that others will have on one's proposals. That is, it becomes harder to ignore implications of prior ideas if one is implicitly aware that one's interlocutors may also be reminded of them by the representations (M. Chi, personal communication, February 6, 1998).

In summary, there is good reason to believe that representational effects will extend to collaborative learning situations in ways worthy of study in their own right.

REPRESENTATIONAL GUIDANCE

Further study is needed because these effects may differ between notational systems and designers of representational tools for collaborative learning need to be informed of the implications of their notational design choices. Representational notations can differ on what information they are capable of expressing (Stenning & Oberlander, 1995), what information they make salient (Larkin & Simon, 1987), and what epistemic processes they promote (Collins & Ferguson, 1993). We claim that the ways in which a collaboratively constructed representational artifact can play the roles just discussed—initiating negotiations of meaning, representational proxy for purposes of deixis, implicitly shared awareness—is sensitive to the notation’s expressiveness and salience of information. Learners use representational tools to construct proxies for their emerging knowledge in a persistent visual medium, inspectable by all participants. The representational guidance of their chosen tool constrains which knowledge can be expressed in the shared context and makes some of that knowledge more salient and hence a likely topic of discussion. Therefore, the cognitive and social affordances of representations might be predicted from the constraints and salience of their notations. This representation-specific influence has been termed *representational bias* (Suthers, 1999a, 1999b) or *representational guidance* (Suthers, 2001). See those studies for detailed predictions, further discussion of the origins of representational guidance, and a comparison of representations in computer-supported collaborative learning (CSCL) systems.

OVERVIEW OF THE STUDY

This article reports on an empirical study of the effects of representational tools on participants’ collaborative discourse and learning outcomes during initial use of those representations in a laboratory setting. In our study, pairs of college science students investigated a problem in the area of public health. They used software based on one of three alternative representational notations (Graph, Matrix, or Text¹) to compile data, hypotheses, and evidential relations, with the goal of coming to a conclusion about the cause of the problem. We measured the influence of

¹Capitalization will be used when referring to these treatment groups or to the specific representations that we provided; lowercase will be used when the generic meanings of “graph,” “matrix,” and “text” are sufficiently precise.

representational tool on participants' activity and talk surrounding evidential relations and on their subsequent elaboration of the data items, hypotheses, and evidential relations that they represent. We also considered learning outcomes as measured by a posttest and post hoc essay. We tested four specific hypotheses.

Our first hypothesis (H1) was that the ontological bias of the representations would affect participants' use of epistemological concepts. We predicted that participants who constructed Graphs or Matrices would classify ideas as "hypothesis" and "data" more than those who constructed text documents. This prediction was made because the Graph and Matrix representations we used *require* that one categorize statements and relations, whereas the Text representation did not have this requirement. To ensure that the effect was due to the representation and not instructions, we were careful to provide similar instructions to all groups and we modeled the classification of information as "hypothesis" or "data" even in the Text conditions. Our prediction was not confirmed.

Our second hypothesis (H2) predicted that participants who constructed Matrices would talk more about evidential relations than participants who constructed Graphs, and that both of these groups would talk more about evidential relations than participants who constructed Text documents. This prediction was made because the representation of evidential relations is no more salient than anything else in a textual representation. In contrast, graphs represent relations with an explicit object (a link) and carry with them the expectation that one construct such links, and matrices prompt for all possible relations with empty fields. The prediction was confirmed for Matrix versus other representations.

Our third hypothesis (H3) predicted that participants who constructed Matrices would elaborate on previously represented information more than those who constructed Graphs, and that both of these groups would elaborate more than those who constructed Text documents. This prediction was made because the notations differ in (a) salience of information (e.g., data and hypotheses are salient in graphs as visual shapes) and (b) whether they suggest consideration of relations between new and previously represented information (e.g., the cells of a matrix prompt for consideration of all relations between row and column items). From a pedagogical standpoint, representations that encourage elaboration of previously represented knowledge help participants integrate that knowledge with their existing knowledge, leading to better retention (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Craik & Lockhart, 1972; Stein & Bransford, 1979). Results confirmed some of the predictions, depending on what was being elaborated. The pattern of results suggests that the quality (if not quantity) of Graph users' elaborations may be highest.

Our fourth hypothesis (H4) predicted that these process differences would lead to significant differences in learning outcomes and subsequent products of the inquiry. Specifically, those who constructed Matrices would remember more data, hypotheses, and evidential relations than those who constructed Graphs, and those who constructed Graphs would remember more data, hypotheses, and evidential

relations than those who constructed text documents. This prediction was made because representations that prompt for increased consideration of evidential relations are in effect prompting participants to elaborate on the information being related. This elaboration in turn should lead to increased memory for the information and greater use of the information in subsequent work. We found no statistically significant differences in recognition memory for information, although the essay contents showed trends in the predicted direction. However, we found that the representational work done by Graph users had a slightly greater impact on the content of their essays than the representational work done by users of Text or Matrix.

In the course of testing these hypotheses and interpreting the results we also measured various other aspects of participants' activity, such as the type and quality of the information they represented. These results call into question whether the quantitative advantages of the Matrix representation translates into quality of discussion, a matter that will require further analyses to resolve.

The remainder of this article is organized as follows. The first two sections present the design of the experiment and the methods of data collection and data coding. The next section reports and interprets our results in five parts:

1. Activities and talk during the learning sessions.
2. Items represented during the learning session.
3. Elaboration of represented items during the learning session.
4. Contents of essays.
5. Posttest results.

The article ends with summary of the results and a discussion of how representational guidance might play out in typical classroom settings. Although this study is limited to initial use of representations in a laboratory setting, we argue that representational guidance of collaborative learning is worthy of further study in both laboratory and natural settings.

DESIGN

We employed a single-factor, between subjects design with three participant groups defined by the software they used: Graph, Matrix, and Text. All three groups were given the identical task of exploring an unsolved science challenge problem—presented as a series of textual web pages—by recording data, hypotheses, and evidential relations as they encountered them. Pairs of participants were randomly assigned to the three treatment groups such that (a) there were no differences in the gender balance of each group and (b) there were no significant differences between the groups' mean grade point averages. Dependent measures included (a) the content of participants' utterances and representational actions and

the timing of these utterances and actions with respect to the availability of information; (b) recognition of the data, hypotheses, and evidential relations explored, as measured by a multiple choice test taken immediately after the learning session; and (c) recall of the most important data, hypotheses, and evidential relations, as measured by the contents of a collaboratively written essay.

Participants

We recruited 60 participants (32 women, 28 men) in self-selected, same-gender pairs out of introductory biology, chemistry, physics, and computer science courses at the University of Hawai'i. Participants were all under 25 years of age and had a mean grade point average of 2.99 on a 4-point scale. All but three participants were native English speakers. The three nonnative speakers were fluent. Participants were paid a \$25 honorarium for their participation.

Materials

Figures 1, 2, and 3 present software used by pairs of participants in the Graph, Matrix, and Text groups, respectively. The left window contains a tool for representing data, hypotheses, and evidential relations. In the Graph version (Figure 1), the left window contains a tool that enables one to build a graph of nodes (data items and hypotheses) and links (evidential relations). To create a data item node (a pink rectangle), one types the data text into the text field centered above the graph drawing area, clicks on the "Add Data" button, and finally drags and drops the node in the graph drawing area. One creates a hypothesis node (a green rounded rectangle) in the same way, except that one clicks on the "Add Hypothesis" button. Finally, to create a link, one clicks on the "Add + link" (supports), "Add - link" (conflicts), or "Add ? Link" (unsure or unspecified) button, and then clicks, in sequence, on the two nodes between which the link is to be positioned.

In the Matrix version (Figure 2), the left window contains a spreadsheet-like tool that enables one to type in data items along the left-hand column and hypotheses along the top row. Clicking on an internal cell of the matrix brings up a pop-up menu with three symbols for evidential relations: "+" (i.e., supports), "-" (conflicts), and "?" (unsure or unspecified), as well as the option of leaving the cell blank. Choosing one of the relation symbols causes the corresponding symbol to appear in the cell, thus relating the data item and hypothesis in the corresponding row and column.

The menu of relations was intended to bring the matrix representation semantically closer to the graph representation and to provide visually salient symbols for the relations. An alternative would be to allow users to fill in the cells however they wished. In a pilot study conducted by Suthers (1999a, 2001), two pairs of secondary school participants were given such a representation (Microsoft Excel pre-

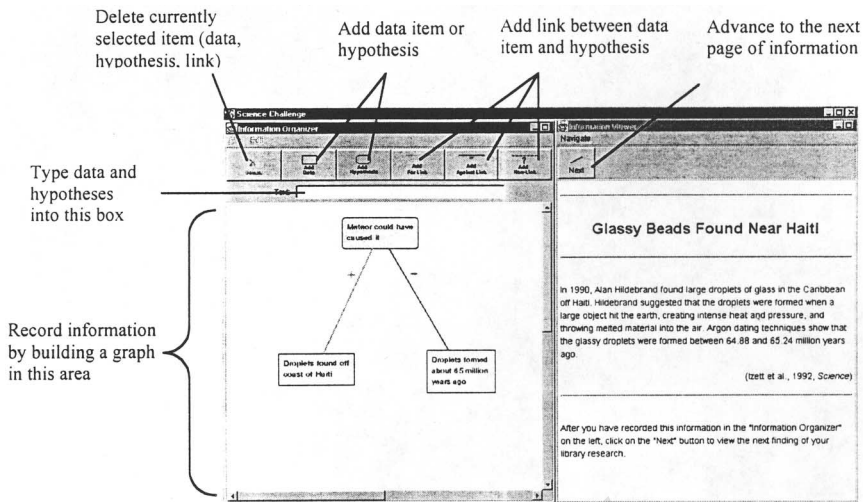


FIGURE 1 The Graph version of the software.

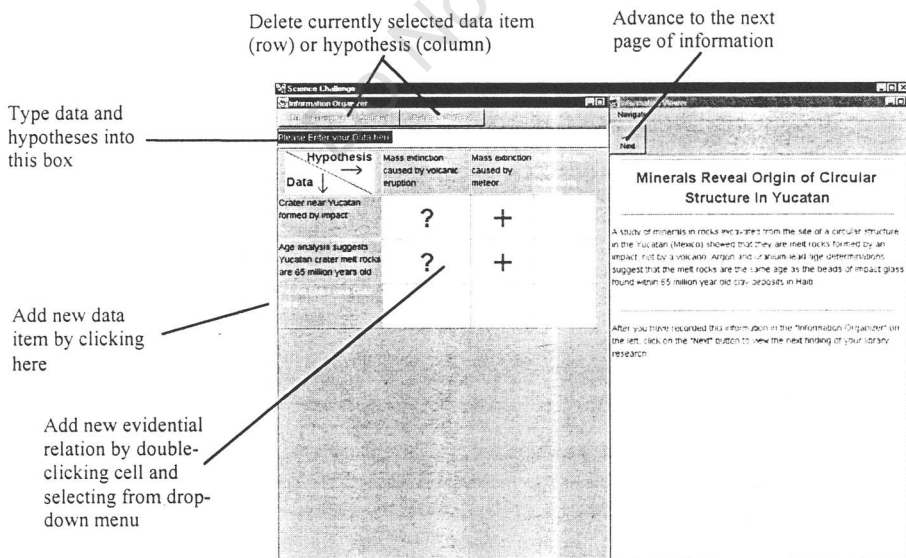


FIGURE 2 The Matrix version of the software.

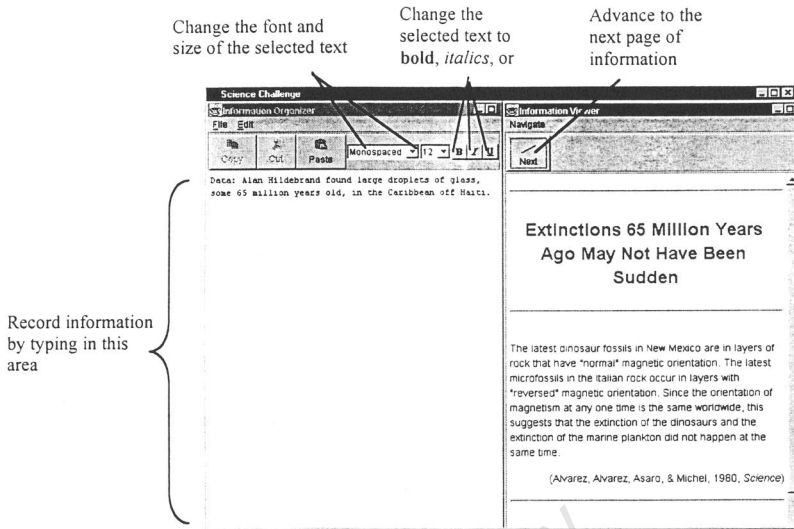


FIGURE 3 The Text version of the software.

pared with large empty cells). Both groups chose to use labels semantically similar to those used here: “supports,” “conflicts,” or “???” in one case and “confirms” or “conflicts” in the other.

The left window of the Text version (Figure 3) contains a simple word processor. Text may be formatted in the usual way—by highlighting the text and then clicking on a formatting button (boldface, italics, underline) or choosing one of several fonts from a list menu.

The right-hand window of all three software versions is identical. This window enables one to advance through a series of textual pages. Each of these pages presents a piece of information pertaining to one of two problems: the cause of mass extinctions at the end of the Cretaceous era or the unsolved mystery of ALS-PD, a neurological disease combining symptoms of Parkinsonism and dementia, which historically had an unusually high occurrence on the island of Guam. One clicks on the “next” button to advance to the next page. The software does not enable the user to revisit previously encountered pages.

We decided to control the sequence of page visitations partly for reasons uncovered during the pilot study (Suthers, 1999a, 2001). In this study, participants freely explored a web of information pages. As a result, different groups viewed different materials, so it was difficult to compare the contents of their work. Also, in this study we wanted to explore the utility of the different representations for relating new information to previous information. We designed a sequence of pages in which new pages bear on the interpretation of information seen several pages earlier. Our analy-

sis relies on the fact that groups viewed the same pages in the same order. A secondary benefit of controlling the sequence of pages is that participants were encouraged to record information in the representational tool as they worked. Although this decision made the task more artificial, the study design was already solidly within an experimental paradigm, so we felt that any further reduction of ecological validity was worth the ability to test for effects sensitive to content sequence.

Task

After some preparation (see Procedure), participants were given a “mission statement” explaining that they were to prepare for an imaginary field trip to Guam by studying some background articles on the ALS–PD disease, with the ultimate goal of discovering the cause of the disease. They would reach this goal by formulating a set of hypotheses regarding the cause of ALS–PD and evaluating data for and against those hypotheses. Participants were instructed that the right-hand window would present background research, one page at a time. They were instructed to record the information on each page using the software tool in the left-hand window and that they would not be able to revisit a page once they clicked on the “next” button. Finally, they were advised that, upon completing all of the pages of background research, they would individually take a short multiple choice test designed to evaluate their familiarity with the information they explored and then work with their partners to write an essay summarizing the results of their research. The essay instructions asked participants to write (a) a brief paragraph describing each hypothesis they formulated and summarizing the evidence for and against the hypothesis and (b) a concluding paragraph that identifies the hypothesis or hypotheses they believe were best supported by the evidence, and justifies this decision. It should be noted that the primary emphasis in the instructions was on uncovering the cause of the disease rather than on preparing for these assessments. Although a laboratory task cannot be considered an *authentic* inquiry task, our review of the videotapes indicates that participants approached their work as a problem-solving task rather than as a memorization task.

Procedure

After an introduction to the study, we provided participants with a brief (10-min) introduction to the software they would be using. The experimenter read aloud and performed a demonstration while participants followed along. Participants then worked on a warm-up science challenge problem (on mass extinctions), which was unrelated to the main problem, so that they could become acquainted with the software and the information-recording process. After 12 min, participants were instructed to stop work on the warm-up problem and to move on to the main problem (ALS–PD). Participants were given as much time as they needed to explore all 15

informational pages on the ALS–PD problem. When they reached the page that informed them that there were no more pages left, the experimenter asked them whether they felt they were done. Some participant pairs decided that they wanted to work further on their representations; they were given as much additional time as they needed. Once a participant pair declared themselves done, the experimenter instructed them to turn off their computer screen, at which point they were given 20 min to individually complete a multiple choice posttest, followed by 30 min to complete a collaborative essay using a word processor.

DATA COLLECTION AND TRANSFORMATION

We had to employ several nontrivial methods to collect and prepare our data for analysis. In this section, we describe the process by which we collected our data, as well as the methods we used to transform our raw data into analyzable data.

Gathering Session Data

We recorded participants' talk in stereo, with the participants' voices recorded in opposite channels. In addition, we recorded two video streams of participants' interaction. A camera positioned behind participants captured their gestures on the screen. A video card output the participants' computer monitor to a second video track. Using a picture-in-picture device, we merged the audio stream with the two video streams such that the behind-the-participants image was inset within the monitor image (Figure 4).

In addition to the video record, we collected automated software logs containing time-stamped records of participants' actions within the software. Using these logs as a starting point, we created detailed transcripts of the 30 participant sessions by transcribing all participant interaction and gestures into the software logs. Participant utterances were broken up into *segments*, based on the principle that a single proposition or idea should occupy a single segment. Likewise, each high-level action in the software (e.g., creating a new data item or hypothesis) was transcribed as a single segment.

Coding Learning Session Segments

We performed a content analysis of participants' learning processes by coding all segments in the 30 transcripts into 8 mutually exclusive *topic* categories:

- *Evidential relation*. These segments consider whether data and hypotheses are consistent or inconsistent; that is, whether a data item supports or conflicts with

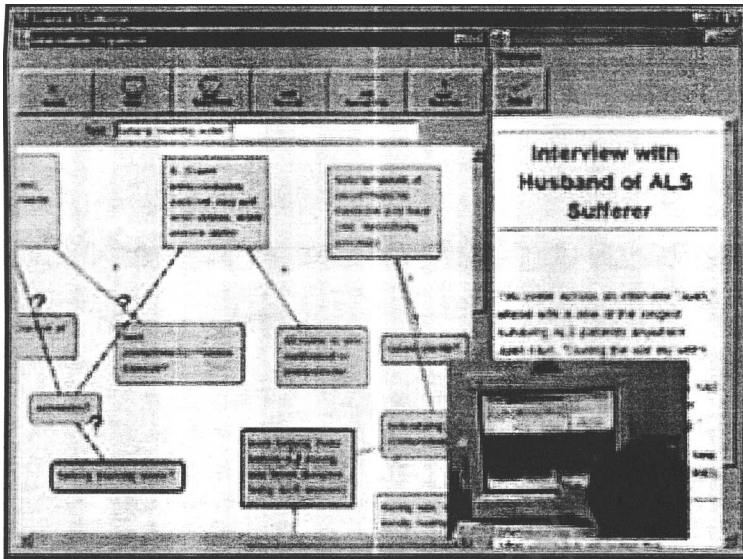


FIGURE 4 Sample of video record of a graph session.

a hypothesis. For example, the segment “That’s for the genetics hypothesis” would be coded as evidential relation—consistency.

- *Epistemic classification.* These segments classify information as either empirical or theoretical; that is, as either data or hypothesis. For example, the segment “Let’s create a new data” would be coded as epistemic classification—data. Likewise, in the Graph software, the action of clicking on the “create data” button would be coded as epistemic classification—data.

- *Reflection.* In these segments, participants step back and either assess what they know so far (e.g., “We know that they used the drinking water for the fadang, to prepare the seeds”), or identify information that is needed but lacking (e.g., “See, but it doesn’t say that these admission records are patients that have the disease”).

- *Warrant.* These segments provide justification for an evidential relation previously cited. For example, the second half of “That supports the aluminum hypothesis, because Irian Jaya was found to have higher than normal levels of aluminum in the soil” would be classified as a warrant.

- *Tool talk.* These segments discuss some aspect of the software. Participants might, for example, ask how to complete some specific task with the software (e.g., “How do you get this out of the way?”), they might complain about the software (e.g., “Oh my, what’s wrong with this thing?”), or they might share their emerging understandings of how the software works (e.g., “If we click on this we can see it”).

- *Domain talk*. These segments discuss the attributes and relations of entities in the domain that participants are exploring. Because our predictions were not specifically concerned with domain talk, it had lower precedence than the aforementioned five categories. For example, “Northern Guam is a low limestone plateau” would be coded as domain talk.

- *On-task*. These segments did not fall into any of the first six categories, but could still be considered on-task. For example, “Let’s go to the next page” would be coded as on-task.

- *Off-task*. These segments were deemed to be unrelated to participants’ learning task. For example, “What did you do last night?” would be considered off-task.

In addition, we coded topic segments with four *modifier* categories, according to whether they were:

- *Verbal* or *representational*—spoken or represented using the software.
- *Recited* or *nonrecited*—quoted verbatim while reading the information pages, or not quoted.
- *Introduced* or *repeated*—the first occurrence of an idea within a given conversation or a reintroduction of an idea already brought up within a given conversation.
- *Conceptual* or *tool-based*—described in conceptual terms (e.g., “It’s supported”) or described with reference to the software (e.g., “Make a plus link”).

To verify the reliability of our coding system, we had two independent analysts code 20% of the transcripts. With respect to the eight mutually exclusive “topic” categories, our analysts attained 89% overall agreement and .86 kappa. With respect to the four modifier categories, agreement levels ranged from 88% (.77 kappa) for introduced versus repeated to 100% agreement (.99 kappa) for verbal versus representational. Given these high levels of agreement, we decided that our coding system was sufficiently reliable, and we then had a single analyst code the remaining 80% of the transcripts.

Coding Information Items

To obtain precise counts of the data items, hypotheses, and evidential relations that participants represented and talked about during their learning sessions, we had to define what should be counted as an “item.” Participants chose to represent and relate the information they encountered in different-sized semantic chunks. For example, upon reading the first information page, one pair created a single data item that read, “Northern Guam is a limestone plateau with high concentrations of calcium in the water.” In contrast, another pair divided the same information into three separate data items: (a) “Northern Guam,” (b) “limestone plateau,” and (c) “high calcium in water.”

To ensure that pairs who chose to divide information into smaller semantic chunks did not get credit for representing or talking about more items, we performed the same task as the participants in our study, identifying a set of 15 data items, 4 hypotheses, and 22 evidential relations that we believe a scientist exploring the materials would have identified.² Table 1 presents a matrix of these data items, hypotheses, and evidential relations. These items, which we call *reference items*, served as normalized semantic units for our counts. Thus, in cases in which participants chose to represent smaller fragments of a given reference item, we collapsed all such fragments into a single item. Participants occasionally created items that were not in our set of reference items. (This happened most frequently in the case of evidential relations.) In these instances, we counted each such item, regardless of its chunk size.

Rather than merely counting items, we also wanted to consider a weighted measure that reflected the fact that (a) certain items were more important to the final analysis of evidence than others and (b) certain items were more difficult to infer than others. Recording data items and hypotheses required few inferences and few judgments of importance; they were the “givens” of the learning task. In contrast, evidential relations required both inference and judgments concerning their relative importance to one’s empirical case. As such, we devised three separate metrics to weight the importance and inferential difficulty of reference evidential relations:

- *Evidential strength*—the strength of the evidential relation, on a scale of 0 to 4 as follows: 0 (*neutral*), 1 (*apparently relevant because it was mentioned in materials*), 2 (*weak correlation or expert opinion*), 3 (*strong correlation or expert opinion*), and 4 (*demonstration of causality*). In Table 1, the modifier + indicates a positive (supporting) relation and – indicates a negative (conflicting) relation (only absolute values were used in the statistical tests).

- *Inferential difficulty*—the number of information pages that must be accessed to infer the relation, with 0 indicating that the relation is explicitly stated in the material, 1 indicating that the relation can be inferred from a single information page, 2 indicating that one must combine information given on two pages, and so on.

- *Inferential span*—the difference (in page numbers) between the first and last page needed to infer the relation. This is a measure of how well participants integrate information given at different pages, which should be sensitive to the utility of the representation.

Weights for our reference evidential relations are included in Table 1. The contents of Table 1 were arrived at by full consensus of both of us (i.e., interrater agreement was 100%).

²The calcium deficiency hypothesis is not a serious contender but was included because it was proposed by many of our participants, perhaps because calcium was mentioned on the first page.

TABLE 1
 “Reference” Data Items, Hypotheses, and Evidential Relations

<i>Data</i>	<i>Hypotheses</i>			
	<i>Calcium</i>	<i>Aluminum</i>	<i>Genetics</i>	<i>Fadang</i>
Northern Guam’s water has been shown to be high in calcium.	+1/1/0			
Southern Guam’s water and soil have been shown to be high in aluminum.	+1/1/0			
All ALS–PD patients in Guam between 1947 and 1952 were native Guamanian.			+2/1/0	
Guamanians eat fadang, the seed of a cycad, and use it as medicine. People regard it as toxic, according to one expert.		+1/3/3		+3,0/0
Guamanians prepare fadang for consumption by soaking it in many changes of water.	+1/3/4	+1/3/3		+1/2/1
Autopsies revealed unusually high concentrations of aluminum in the brain tissue of ALS–PD patients—300–600 p.p.m., rather than 1–3 p.p.m.		+3/2/4		
Umatac, on Guam’s southwest shore, had by far the highest rate of ALS–PD in the 1950s.	–2/2/6	+2/2/5	+3/2/2	
Navy doctor observed several related patients in Guam hospital with full-blown ALS–PD; he suspected the disease could be genetic.			+2/0/0	
Umatac is geographically and genetically isolated, made up almost entirely of people from four interrelated families.			+3/2/2	
A high frequency of ALS–PD cases has been reported in Irian Jaya, where residents use cycads for medicine, and where the water and soil have much higher than normal concentrations of aluminum.		+3/2/8	–2/2/2	+2 /2/6
Aluminum is the third most abundant element on earth and many places have aluminum-rich soil. One expert asserts that ALS–PD in Guam cannot be caused simply by exposure to aluminum.		–3/0/0		
Monkeys that were fed BMAA (found in cycad seeds) exhibited ALS–PD-like symptoms.				+4/2/8
BOAA, a natural toxin with a chemical structure similar to BMAA, has been found to cause an ALS–PD-like disease called lathyrism.				+3/3/9
The outstanding pathological feature of ALS–PD is neurofibrillary tangles. The brains of animals injected with BMAA did not have this feature.				–4/3/10
In studying the family histories of Umatac, a geneticist found no definitive pattern typical of a genetic disease.			–4/3/8	

Note. Each reference evidential relation is identified by its polarity (+: supports, –: conflicts with) and three numbers: evidential strength, inferential difficulty, and inferential span. See text for explanation.

Coding Elaborations

Several of our analyses are concerned with the extent to which alternative representations encourage participants to revisit and elaborate on previously represented items. Because we were specifically interested in elaborations that may have been prompted by participants' representations rather than by our information pages, our analysis operationalized *elaboration* as any reference to an item that took place while participants were viewing an information page that followed the page they were viewing when they initially represented the item. A reference could take any of the following four forms:

- An explicit verbal reference to the item.
- An implicit verbal reference to the item through the item's representational proxy.
- A verbal or representational formulation of, or reference to, an evidential relation that includes the item (in the case of data items and hypotheses).
- A representational change (e.g., changing an evidential relation from "+" to "-" or changing the wording of an item).

Scoring Posttests

Each of 13 multiple choice test questions had four possible answers plus "none of the above." We designed the posttest such that some questions had more than one answer and developed the following scheme for awarding partial credit. We gave one point for each correctly circled answer and we subtracted half a point for each incorrect answer, subject to the constraint that it was not possible to score below 0 on any given question.

Coding and Scoring Essays

Finally, we performed counts of the data items, hypotheses, and evidential relations that participants included in their collaborative essays. Just as we did with the session transcripts, we counted each item or group of items that sufficiently matched a reference item and also counted each "custom" item that fell outside the scope of our reference items. Once we had obtained essay item counts, we scored each reference evidential relation cited by participants according to their evidential strength, inferential difficulty, and inferential span—the same three measures we used to weight items represented during the learning session.

RESULTS AND DISCUSSION

Our presentation and discussion of results is organized into three sections concerned with the learning session followed by two sections examining outcome

measures. Table 2 provides an advance organizer. Summaries of the results are indexed by the sections (titled by data category) and subsections (titled by measures) in which they are discussed. Nonsignificant results are in italics. The table also indicates the hypothesis (from the introduction) that is most directly tested by each measure and the tables in which the relevant data are reported.

Learning Session: Activities and Talk

We begin by reporting baseline measures of participants' activities and talk and then test our first two hypotheses.

Baseline counts. Table 3 presents the mean time that participants in the three conditions took to complete the learning session. On average, the Text group finished over 6 min faster than the Graph group, and over 8 min faster than the Matrix group; however, an analysis of variance (ANOVA) found no significant differences between the groups ($df = 2, F = 1.38, p < .27$).³

Table 4 lists the mean number of total On-Task, Nonrecited segments by condition; these totals are also broken down into Verbal and Representational segments. There was no significant difference between treatment groups on the total counts ($df = 2, F = 1.02, p < .3748$), nor on Verbal counts ($df = 2, F = .90, p < .4168$) or Representational counts ($df = 2, F = 1.17, p < .3260$).

To provide an overview of the high-level content of these segments, Figure 5 graphically presents the percentage of these segments dedicated to each of the high-level topic categories described in a previous section. As the figure suggests, all three treatments dedicated roughly two thirds of their sessions to talk and activities focusing on domain and task-oriented issues. In all cases, a small number of segments (between 1% and 3%) focused on the software itself, whereas equally small numbers of segments focused on higher order thinking in the Reflection and Warrant categories.

Epistemological classifications. Our first hypothesis (H1) predicted that the explicit use of the concepts of "hypothesis" and "data" in the Graph and Matrix interfaces would lead participants to explicitly classify information using these terms more than in the Text condition, where no such representational prompting

³The statistical significance measures reported in this article employ two different sets of statistical methods. In cases in which we are comparing relatively continuous numeric data (e.g., means of posttest scores or counts), we use an analysis of variance (ANOVA) and a standard post hoc parametric test (Tukey). In contrast, in cases in which we are comparing ratios with varying denominators (which we report as percentages), the assumptions of the parametric ANOVA are violated because the data are not of interval scale. Therefore, in these cases, we use a nonparametric Kruskal–Wallis test, along with a nonparametric post hoc Fisher–Protected Least Significant Difference test.

TABLE 2
Summary of Results Indexed by Sections in Which They are Discussed

<i>Data Category</i>	<i>Measure(s)</i>	<i>Hypothesis</i>	<i>Tables</i>	<i>Summary of Results</i>
Learning session				
Activities and talk	Baseline counts		3, 4	<i>No significant differences in time on task or overall quantity of activity or talk.</i>
	Epistemological classifications	H1		<i>No significant differences in use of theoretical and empirical categories.</i>
	Evidential relations	H2	5	Matrix users were more concerned with evidence.
Representation of items	Baseline counts	H2	6, 7	Several differences exist in quantity and coverage of data, hypothesis, and evidential relations represented.
	Relating represented data and hypotheses	H2	8	Matrix users represented more of the available evidential relations.
Elaboration of represented items	Revisitation of data and hypotheses	H3	9, 10	Elaboration was greater in structured representations than in Text. No temporal differences were found.
	Revisitation of evidential relations	H3	11	Elaboration was low in all groups, but greater in Matrix due to repeated revisitation and modification of relations.
Outcome measures				
Essays	Baseline counts		12	Text users wrote about more hypotheses than the other groups.
	Representation-essay overlap	H4	13	The work done with representations has the greatest influence on essay content in Graph.
	Inferential quality of evidential relations	H4	14	<i>Trends in predicted direction were not significant.</i>
Posttest	Posttest scores	H4	15	<i>No differences were found.</i>

TABLE 3
Mean Time to Complete Learning Task,
by Treatment, in Minutes and Seconds

<i>Treatment</i>	<i>Time on Task</i>	
	<i>M</i>	<i>SD</i>
Matrix	44:15	15:19
Graph	46:51	12:13
Text	38:04	7:36

TABLE 4
Mean Number of Total, Verbal, and Representational
Segments That Were On-Task and Nonrecited

<i>Treatment</i>	<i>Total</i>		<i>Verbal</i>		<i>Representational</i>	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Matrix	510.0	163.1	396.2	132.6	113.8	61.2
Graph	454.5	232.3	365.8	195.0	88.7	38.8
Text	392.5	145.8	305.0	126.1	87.5	20.3

was present. However, the numbers do not bear this out. A nonparametric Kruskal–Wallis test⁴ does not detect a significant difference ($df = 2, H = 4.42, p = .1099$). We speculate that Text users were just as likely to classify information as either “hypothesis” or “data” because our demonstration of the Text software deliberately used these terms as labels, in order not to bias the results with different instructions. Apparently, participants faithfully followed our example.

Evidential relations. Our second hypothesis (H2) predicted that Matrix users would focus on evidential relations more than Graph users, and that both groups would do so more than Text users. Figure 5 shows a notable difference in the predicted direction. We present the mean number of segments concerned with evidential relations vis-à-vis the three conditions in Table 5. This measure is given both as raw counts and as percentages of the total on-task, nonrecited segments. These counts and percentages are further broken down according to whether they are representational (i.e., actions leading to representation changes in the software tool used) or verbal (i.e., spoken).

To test for differences between group percentages, we used a nonparametric Kruskal–Wallis test. Beginning with the evidential relations, we found significant

⁴See Footnote 3 for an explanation of our choice of statistical tests.

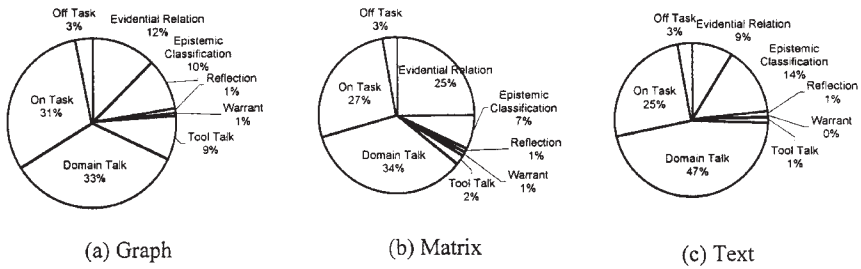


FIGURE 5 Mean percentage of On-Task, Nonrecited segments that Graph, Matrix, and Text participants dedicated to the eight high-level topic categories.

differences with respect to overall percentages of evidential relation segments ($df=2, H=8.712, p < .013$) and with respect to the percentages of verbal evidential relations ($df=2, H=12.56, p < .0019$). (Statistical comparisons of verbal events were performed on percentages of total *verbal* segments, not percentages of total segments.) A post hoc Fisher PLSD test determined that, in both cases, the significant differences were between Matrix and Graph ($p < .05$) and between Matrix and Text ($p < .05$).

These results furnish evidence for our general hypothesis that the type of representations that learners use in collaborative investigations will impact the focus of their discourse (H2). In particular, discussion of relations can be influenced by the extent to which a representational toolkit used by learners prompts for consideration of those relations. This effect was obtained when considering all propositional acts (whether expressed verbally or by a modification to the representation) and when considering only verbally expressed propositions. However, differences in the use of concepts predicted from representational considerations alone failed

TABLE 5
Mean Evidential Relations Segments as Counts and as Percentages of the Total, Verbal, and Representational On-Task Segments

Treatment	Total		Verbal				Representational					
	Count		%		Count		%		Count		%	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Matrix	139.2	97.0	27.3	13.4	67.4	60.7	59.2	19.9	71.8	44.9	18.1	9.3
Graph	60.3	19.1	15.1	6.3	25.6	10.2	29.9	7.3	34.7	12.3	11.3	6.1
Text	37.7	27.2	9.6	6.3	15.4	7.9	17.7	11.1	22.2	21.7	7.3	4.8

to materialize. Instructions coupled with behavioral modeling (i.e., a demonstration) can sometimes make up for lack of representational prompting. Yet instructions are not always the dominant factor. Our demonstration of the Text software also included an explicit demonstration of how to record an evidential relation, yet the data still support the predicted differences in learners' focus on such relations.

Learning Session: Representation of Items

Using the software provided, participants represented numerous data items, hypotheses, and evidential relations as they visited the trail of web pages they encountered during their learning sessions. In this section, we present three separate analyses that focus on the items represented by participants.

Baseline counts. Table 6 provides a baseline for the analyses of this section by summarizing the data items, hypotheses, and evidential relations that participants in each treatment group represented, measured by both mean counts and percentages of our reference items.

We note several trends in these data. First, although the three groups were identical in terms of number of represented data items (first numeric column of Table 6), the Text group represented more hypotheses than the other two groups, as reflected in the counts and percentages of reference hypotheses represented (5th and 7th numeric columns of Table 6). An ANOVA on the counts indicates that this difference is statistically significant ($df = 2, F = 4.80, p = .0165$); a post hoc Tukey test reveals that the difference is between Text and Graph ($p < .05$).

Second, the Matrix group represented substantially more evidential relations than the other two groups, as reflected in the counts and percentages of reference evidential relations represented (9th and 11th numeric columns of Table 6). This difference, according to an ANOVA on the counts, is statistically significant ($df = 2, F = 7.21, p = .031$), with the differences lying between both Matrix and Graph

TABLE 6
Data Items, Hypotheses, and Evidential Relations Represented in
Learning Session, Both as Mean Counts (Including Nonreference Items)
and as Mean Percentages of Reference Items

Treatment	Data				Hypotheses				Evidential Relations			
	Count		%		Count		%		Count		%	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Matrix	14.8	0.6	98.7	4.2	5.3	3.0	72.5	7.9	47.5	40.2	63.2	22.8
Graph	14.7	0.5	98.0	3.2	3.8	1.4	57.5	16.9	9.2	5.1	25.0	18.0
Text	14.7	0.7	98.0	4.5	7.2	2.7	80.0	23.0	15.0	11.4	30.9	20.1

(Tukey test, $p < .05$) and between Matrix and Text (Tukey test, $p < .05$). This result echoes our prior results concerning discussion of evidence and provides indirect support for H2.

Third, the large difference in number of evidential relations represented translates into a statistically significant difference in total number of items represented ($df = 2$, $F = 6.68$, $p = .0044$; see Table 7). Post hoc Tukey tests show the difference to be between both Matrix and Graph ($p < .05$) and between Matrix and Text ($p < .05$).

Finally, we consider the mean counts in relation to our reference items. On average, participants represented 14.7 data items, 98% of our 15 reference items. This indicates that participants are in high agreement with us concerning the 15 data items to be gleaned from the materials. In contrast, Matrix and Text had on average more hypotheses than we did (5.3 and 7.2 compared to 4) and Matrix had far more evidential relations (47.5 compared to 22). Clearly, Matrix users were not as discriminating as we were in creating evidential relations.

Relating represented data and hypotheses. Matrix participants may have represented significantly more reference relations than the participants in the other two groups merely because the data and hypotheses to be related were not available in the representations of the other groups, and not because of prompting by the Matrix notation. To rule out this explanation, we consider the extent to which participants represented relevant evidential relations upon representing the corresponding data item and hypothesis to be related. To carry out such an analysis, we focus on our set of 22 reference evidential relations because these are the only relations that we can reasonably expect participants to represent. Focusing on only those evidential relations for which both relata were previously represented, Table 8 lists the mean percentage of those reference evidential relations that were actually filled in by participants across treatment groups. A nonparametric

TABLE 7
Total Items Represented in Learning Session, Both as
Mean Counts (Including Nonreference Items) and as
Mean Percentages of Reference Items

Treatment	Total Items			
	Count		%	
	M	SD	M	SD
Matrix	67.6	41.8	77.1	12.4
Graph	27.7	5.6	54.9	11.3
Text	36.9	13.3	60.2	12.0

TABLE 8
 Mean Percentage of Available Reference
 Evidential Relations That Were
 Represented

<i>Treatment</i>	% Represented	
	<i>M</i>	<i>SD</i>
Matrix	72.5	25.8
Graph	33.2	21.8
Text	34.2	23.6

Note. By available, we mean evidential relations whose data and hypothesis components had already been represented.

Kruskall–Wallis test detects a statistically significant difference in these percentages ($df = 2$, $H = 11.19$, $p = .0037$). Post hoc Fisher PLSD tests show that the difference is between Matrix and Graph ($p < .05$) and between Matrix and Text ($p < .05$).

These results are consistent with our reasoning that Matrix users filled in significantly more evidential relations because of a property of the notation (H2). It appears that the empty cells created when one represents a new data or hypothesis in the Matrix prompt users to fill in the available evidential relations. In the other two representations, available evidential relations are perhaps not as obvious, so one is less likely to attend to them.

Learning Session: Elaboration of Represented Items

We now turn our attention to a question closely related to the one just considered: To what extent do the alternative representations encourage participants to *elaborate on* previously represented items? This section presents two analyses that explore this question.

Revisitation of data and hypotheses. First, we consider the extent to which participants revisited, within their learning sessions, the data and hypotheses that they initially represented. (See the section called “Coding Elaborations” for a precise definition of *revisitation* as a proxy for elaboration.) We hypothesized that the Matrix group would revisit represented data and hypotheses more than the Graph group and that the Graph group would revisit represented data and hypotheses more than the Text group (H3). Our reasoning was that the Matrix representation encourages participants to reconsider represented data and hypotheses because they are being prompted by cells to be filled in to explore all possible evidential relations. In contrast, because the Text representation does not make data and hypotheses salient as visual objects and does not explicitly represent evidential relations, we reasoned that it would not prompt participants to reconsider

the data and hypotheses that they write down. We predicted that the Graph representation would lie somewhere in the middle of these two representations. The Graph representation should encourage elaboration because data and hypothesis statements are reified as visual objects (shapes) arranged on the screen. The salience of these objects was expected to encourage subsequent discussion of the corresponding statements through reminding and ease of deictic reference. However, revisitations would be less frequent than in Matrix, because although Graph explicitly represents evidential relations by links, it does not explicitly represent their absence, so it does not encourage exploration of all possible relations.

Table 9 presents the mean ratios and percentages of represented data and hypotheses that participants revisited in their learning sessions. (The denominators of the ratios are the sums of the counts of represented data and hypothesis items from Table 6.) As these numbers indicate, there exists a gap between both the Graph and Text groups and between the Matrix and Text groups. A nonparametric Kruskal–Wallis test of the mean percentages indicates that there does indeed exist a statistically significant difference ($df = 2$, $H = 10.21$, $p = .0061$) and post hoc Fisher PLSD tests confirm that the difference lies between both Graph and Text ($p < .05$) and between Matrix and Text ($p < .05$). These results confirm our hypothesis that Matrix and Graph are superior to Text for prompting elaboration on represented information (H3). Contrary to our prediction that Matrix would elaborate more than Graph, the numbers for Graph are higher than Matrix, but this difference is not significant. The absence of the predicted difference is interesting in light of the fact that Matrix users represented significantly more relations, which increases elaboration counts for the data and hypotheses being connected. Apparently, Graph users achieved equally high levels of elaboration on data and hypotheses for

TABLE 9
Mean Ratios and Percentages of Represented Data Items
and Hypotheses Revisited Within the Learning Session

<i>Treatment</i>	<i>Ratio</i>		<i>%</i>	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Matrix	12.2	4.0	61.6	21.5
	20.1	3.2		
Graph	13.3	3.7	71.9	18.8
	18.5	1.4		
Text	8.6	4.3	39.3	18.8
	21.9	2.7		

Note. The denominators of the ratios are the sums of the counts of represented data and hypothesis items from Table 6.

reasons other than representing many relations between them, suggesting that direct elaboration on data and hypotheses is supported by the visual salience of their representational proxies (a question to be explored in future analyses).

Having detected general trends, we now turn to an analysis of the distribution and timing of reintroduction events. Does the number of revisitations per item differ between representations? Do participants tend to revisit items fairly recently after they represent them, or much later in the session, perhaps as their relevance becomes evident to a discussion?

We answered these questions by examining logs of revisitation events indexed by (a) the number in sequence of the information page that was visible when the event occurred (there were 15 total information pages) and (b) the number of the segment in which the event occurred. Summary data from these logs are presented in Table 10. A “span” is the number of pages or segments from the initial representation of the item to the revisitation event.

A nonparametric Kruskal–Wallis test did not yield significant differences between the groups with respect to the mean number of revisitations per item ($df = 2$, $H = 5.23$, $p = .0732$). However, trends in both means and standard deviations suggest that Matrix pairs are revisiting a few items more often than in the other groups. Although we previously saw that the Graph pairs revisited slightly more data and hypotheses overall (Table 9), they did not revisit each of those items as often as did Matrix pairs.

With respect to the average page and segment span of each revisitation, a nonparametric Kruskal–Wallis test detects no significant differences (page span: $df = 2$, $H = 4.51$, $p = .1047$; segment span: $df = 2$, $F = 3.42$, $p = .1805$). We speculate that this lack of difference indicates that the sequencing of information in the pages has more impact on the *timing* of revisitations than does representation type.

Revisitation of evidential relations. We now consider participants’ revisitation of previously represented evidential relations. Table 11 presents mean ratios and percentages of revisited evidential relations. A comparison to Table 9 indicates that subsequent elaboration of evidential relations was much more rare than elabo-

TABLE 10
Mean Number of Revisitations per Data Item/Hypotheses
and the Mean Page and Segment Spans per Revisitation

Treatment	Revisitations Per Item		Page Span Per Revisitation		Segment Span Per Revisitation	
	M	SD	M	SD	M	SD
Matrix	4.7	4.2	6.0	1.3	326.2	114.2
Graph	1.7	0.3	5.2	1.7	275.0	203.7
Text	2.7	1.5	4.9	0.6	224.4	83.8

TABLE 11
 Mean Ratios and Percentages of Evidential
 Relations Revisited Within the Learning Session

<i>Treatment</i>	<i>Ratio</i>		<i>%</i>	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Matrix	$\frac{7.3}{47.5}$	$\frac{9.2}{40.2}$	14.8	20.4
Graph	$\frac{0.2}{9.2}$	$\frac{0.4}{5.1}$	2.1	4.4
Text	$\frac{0.8}{15.0}$	$\frac{1.1}{11.4}$	5.0	7.3

ration of data and hypotheses. However, despite this lower frequency, a nonparametric Kruskal–Wallis test of the groups’ mean percentage of revisited evidential relations yields a significant difference between the groups ($df = 2$, $F = 6.85$, $p = .0325$). A post hoc Fisher PLSD test shows the difference to be between Matrix and Graph ($p < .05$).

To explain the fact that participants revisited evidential relations less frequently than they revisited data and hypotheses, we observe that evidential relations are already a synthesis of the information that participants encountered. Indeed, representing an evidential relation constitutes a more reflective activity than representing a data or hypothesis. Therefore participants in our study may not see evidential relations as requiring further reflection. Revisitation of evidential relations may be more common in other task domains, such as legal argumentation, where much of the discourse focuses on the warrants behind one’s inferences.

We have two explanations for the differences in revisitations between Matrix and Graph, both of which suggest problems with these representations. First, although 46% of the revisitations of relations in Matrix were changes to the type of relation, there was only one change event in all of the Graph sessions and none in Text. Therefore, we believe that Matrix users felt compelled to modify their relations much more often than other participants because they were prompted by the cells to invent relations between items that were not particularly relevant to each other (as well as items that were). The video data includes many examples of Matrix participants changing each relation several times while they attempted to resolve the ambiguity. This interpretation is consistent with the high number of relations (and therefore of nonreference relations) indicated by Table 6.

The second explanation requires understanding relevant details of the software tools. In the Graph tool, one creates a new relation by selecting the appropriate relation’s icon (“+,” “-,” or “?”) and then selecting the statements that form the start and

endpoint of the link in turn. The method of changing the type of an existing link is entirely different: one must either right-click to obtain a link editor or delete and then recreate the link. In contrast, the method of changing a relation in Matrix is identical to the method of creating it in the first place: One selects the cell of the Matrix to obtain a menu of options. We speculate that there would be more revisitations in Graph if the method of modifying the relation became obvious while creating it. This discussion illustrates the importance of considering one's instructional objectives even in the design of the most basic human-computer interactions.

Outcome Measure: Essays

Recall that participants wrote collaborative essays roughly 25 min after they completed their learning sessions. We now consider the extent to which participants included in their essays those items that they represented during their learning sessions. Do participants tend to remember and integrate into their own findings those items that they represented during the learning session? Does the impact of the representational work differ between treatment groups?

Consider these questions with respect to two null hypotheses. The most basic null hypothesis is that there is no relation between representation and essay content, and therefore will be no content differences between essays. Alternatively, we might hypothesize that there is a relation, but that it is independent of the particular representation being used: Representing an item increases the likelihood that it will be remembered and included in the essay, regardless of the representation. If this were indeed the case, we would expect the Text group to include significantly more hypotheses in their essays than the Graph group and the Matrix group to include significantly more evidential relations in their essays than the other two groups because this is the pattern of representational counts found in Table 6. Departures from this pattern may admit the possibility that the relation between representation and subsequent use of information may be specific to the type of representation used (H4). The following subsections explore these hypotheses.

Baseline counts. Table 12 presents the mean number of data, hypotheses, and evidential relations participants included in their essays. An ANOVA indicates that the Text group included significantly more hypotheses than the Graph group ($df = 2, F = 4.79, p = .0166$; Tukey test: $p < .05$); however, there exist no differences between the groups with respect to number of evidential relations included in their essays ($df = 2, F = .19, p = .8318$). Thus, there are content differences between essays and we may reject the first null hypothesis. Also, more items represented in the session did not necessarily translate into more items discussed in the essay, so we may also reject the second null hypothesis.

TABLE 12
Data Items, Hypotheses, and Evidential Relations Represented in
Essays, as Mean Counts and Mean Percentages of Reference Items

Treatment	Data				Hypotheses				Evidential Relations			
	Count		%		Count		%		Count		%	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Matrix	10.6	3.0	65.3	22.4	4.8	1.1	65.0	24.2	11.2	4.1	37.3	20.3
Graph	9.8	3.2	62.0	16.6	3.7	1.3	60.0	24.2	10.1	5.9	36.8	21.7
Text	10.5	3.4	63.3	20.7	5.3	1.1	75.0	16.7	9.9	5.3	30.9	16.0

Representation-essay overlap. The fourth representational guidance hypothesis (H4) predicts that there will be group differences with respect to the percentage of *carryover* items: those data items, hypotheses, and evidential relations that were represented in the session and subsequently included in the essay. To test this hypothesis more directly, we computed each group's percentage of represented-in-session items that were also included in the essay (see Table 13). Inspecting these percentages, we find that the Graph condition had a higher percentage of carryover items than both Matrix and Text. A nonparametric Kruskal–Wallis test indicates that this difference is statistically significant ($df=2, H=6.48, p=.0391$). A post hoc Fisher PLSD test shows that the difference is between Graph and Matrix ($p < .05$).

In interpreting this result, we note that Graph users were more focused with respect to what they represented in their learning sessions. Table 6 tells us that they were more selective than users of other representations in both the hypotheses and the evidential relations that they represented. Graphs prompt users to identify and represent *some* relation involving each new item, but does not specify which relation, and (unlike a matrix) does not encourage representation of all possible relations. Thus pairs are faced with the need to discuss which relation to represent so they engage in a discussion of the possible relations and their significance. We therefore speculate that Graph pairs are encouraged to engage in higher order

TABLE 13
Mean Percentage of Represented-in-Session Items Included in Essays

Treatment	Total		Data		Hypotheses		Evidential Relations	
	M	SD	M	SD	M	SD	M	SD
Matrix	36.2	21.0	66.2	22.1	80.4	28.5	20.9	23.0
Graph	55.4	17.6	63.1	16.0	71.8	25.1	36.4	33.1
Text	48.9	16.3	64.3	20.2	56.1	14.7	35.4	29.9

thinking when faced with the choice of how to connect a newly added item. (This speculation will be tested in a forthcoming discourse analysis.)

In contrast, Text users (who had the most hypotheses in both the session representations and the essays, yet the least overlap between the two) were less discriminating in the hypotheses they represented and were not prompted to evaluate these hypotheses in any particular way, so apparently they reinvented hypotheses as they wrote their essays. (There is a nonsignificant trend for hypotheses according to a Kruskal–Wallis test, $df = 2$, $H = 5.27$, $p = .0716$; 46% of the hypotheses in Text essays were new.) Matrix pairs may have filled in cells (47.5 cells on average) without discriminating which relations were important, but then selected a smaller set of relations (11.2 on average) while writing their essays. This interpretation is corroborated by our analyst’s informal observation that some Matrix groups filled in the cells late in the session by systematically going down columns or across rows with minimal discussion, whereas Graph users usually linked items as they went, discussing each link.

Inferential quality of evidential relations. Finally, we note disappointing results concerning one measure of the quality of inferences in the essays. Weighting evidential relations included in the essays by their evidential strength, inferential difficulty, and inferential span, just as we did in the case of evidential relations represented during the learning session, we obtain Table 14. ANOVAs reveal no significant differences between the groups’ evidential strength ($df = 2$, $F = 1.03$, $p = .3718$), inferential difficulty ($df = 2$, $F = .98$, $p = .3883$), and inferential spread scores ($df = 2$, $F = .64$, $p = .5327$), although trends for all three are largely in the predicted direction (Text < Graph < Matrix).

Outcome Measure: Posttest

We administered a multiple choice posttest of domain knowledge to individuals, but averaged the scores of each pair’s members for purposes of analysis because the group is the unit of analysis in this study. We hypothesized that pairs who represented more items, represented more important items, and elaborated more on the

TABLE 14
Mean Evidential Strength, Inferential Difficulty, and Inferential Span Scores for Evidential Relations Included in Essay

<i>Treatment</i>	<i>Evidential Strength</i>		<i>Inferential Difficulty</i>		<i>Inferential Span</i>	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Matrix	2.17	0.58	1.40	0.49	3.38	1.23
Graph	2.26	1.39	1.37	0.81	3.12	1.97
Text	1.66	0.91	1.04	0.55	2.56	1.67

items that they represented would remember those items better, leading to higher scores on the posttest (H4). This hypothesis translates into a prediction for higher posttest scores for the Graph group, as compared to the Text group, and for the Matrix group, as compared to the other two groups.

Table 15 presents the average scores on the posttest. An ANOVA yielded no significant differences between the three groups ($df = 2, F = .046, p < .96$); in fact, all groups obtained approximately 68% to 69% correct. We can offer three possible explanations for the lack of learning outcomes despite the fact that we have just reported significant differences in *processes* such as elaboration that other literature suggests should lead to improvements in learning outcomes. Our posttest, which contained 13 multiple choice questions, may not have been sensitive enough to detect actual learning differences between the groups. Also, participants did not spend enough time in the learning session (approximately 45 min; see Table 3) for the learning outcomes to develop fully. Alternatively, measurable differences in retention may not emerge until more substantial time has passed.

GENERAL DISCUSSION

Motivated by theoretical considerations (Suthers, 1999b, 2001), informal studies (Suthers & Weiner, 1995), and classroom trials (Suthers et al., 1997; Toth et al., 2002), this study investigated representational effects on the learning processes and outcomes of collaborating learners. Participants were provided with one of three representational notations for recording information relevant to deciding between alternative explanations of a public health problem: Graph, Matrix, or Text. Dependent measures included (a) the content of participants' utterances and representational actions and the timing of these utterances and actions with respect to the availability of information; (b) a multiple choice test of the ability to recognize the data, hypotheses, and evidential relations explored; and (c) the contents of a written essay summarizing the inquiry. Our key results included the following:

TABLE 15
Mean Posttest Scores

<i>Treatment</i>	<i>Posttest Score</i>	
	<i>M</i>	<i>SD</i>
Matrix	15.2	3.5
Graph	15.2	3.4
Text	14.9	3.5

Note. Scores are out of 22.

- Graph users represented the fewest items. Text and Matrix users represented more hypotheses than we derived from the materials and Matrix users represented far more evidential relations than were in our analysis (H2, see Tables 6 and 7).
- The exhaustive prompting of Matrix for consideration of all possible evidential relations led participants to discuss and represent issues of evidence more than users of other representations (H2, see Tables 5 and 8).
- Users of visually structured representations (Graph, Matrix) revisited previously discussed ideas more often than users of Text (H3). Matrix users revisited prior data and hypotheses mainly to fill in the matrix cells that relate them.
- Revisitation of relations was rare except for Matrix users, who often modified their relations (H3, see Table 11).
- The representational work done by Graph users had a slightly greater impact on the content of their essays than the representational work done by users of Text or Matrix (H4, see Table 13).

We take these results as an existence proof that the choice of representational notation for collaborative learning applications can have significant effects on learner's interactions (H2 and H3) and may differ in their influence on subsequent collaborative use of the knowledge being manipulated (H4). Specifically, visually structured and constrained representations can provide guidance for collaborative learning that is not afforded by plain text. However, not all guidance is equal and more is not necessarily better. For example, it is possible to overprompt for consideration of irrelevant relations. Whether a given representation's prompting is desirable is a pedagogical decision that must be considered in light of the educational objectives and context of the activity.

However, the effects described here should not be understood as deterministic. Experts benefit from a representation's affordances without being misled by them. Appropriate instruction and experience can help learners develop similar sophistication in the use of representations. We believe that each representation has its own strengths and weaknesses, and each may be the best choice for a given application. In fact, our current version of Belvedere⁵ integrates three representational "views" of evidence models (Graph, Matrix, and a Hierarchy representation not discussed here) in one tool, providing an interesting platform for future studies. We speculate that Graph will be most useful for gathering and relating information by the relations that motivated its inclusion and Matrix will be most useful for subsequently checking that no important relations have been missed (particularly in domains exhibiting multiple causality) and for scanning for patterns of evidence.

⁵See <http://belvedere.sourceforge.net/> or <http://lilt.ics.hawaii.edu/lilt/software/belvedere/>

EDUCATIONAL IMPLICATIONS

What significance do these results from a controlled laboratory study have for collaborative use of representations in typical school settings? In this section we discuss ways in which our results might extrapolate to school settings and then describe a related classroom study that sheds some light on the matter.

The classroom environment differs from the laboratory in many ways, perhaps the most significant being that there would be many groups of students working on their projects and examining each other's work and a teacher circulating to supervise and guide these activities. As a result, there would be additional conversations centered on the representations. We might expect representational notations to influence these conversations as well. Teachers, as well as peers visiting from other groups, may not have participated in the negotiations that infuse representations with meanings for those who have built them. Nonetheless, a representation that makes the conceptual content and potential for further constructive work salient would better support cross-group and teacher coaching than one that did not. For this reason, we have designed the present version of Belvedere to facilitate quick overviews (e.g., by visually marking links when a justification annotation has been added to the link).

Other differences to be expected in the classroom are (a) a higher student-computer ratio, (b) less restricted access to information, and possibly (c) greater time on task.

Larger group size can result in social effects such as decreased involvement of some of the individuals (Slavin, 1990; Webb & Palincsar, 1996). Although these problems arise for reasons unrelated to the representations, visually salient representations may enable those not close to the screen to track and comment on activity better than textual representations.

Representational effects were suggested by our pilot study (Suthers, 1999a, 2001), which did not restrict or control access to information. However, some students were inclined to search through much of the materials before beginning to use the evidence modeling tool provided. Under similar circumstances in the classroom, representational guidance may not be a factor until the students started using the tool.

Increasing time on task in the classroom requires curricular changes and would benefit from block scheduling. With greater time on task we would want to examine how learners' use of representations changes over time. In what ways would a community of learners spontaneously develop shared semantics for a given representational notation and strategies for their use? How would these semantics and strategies spread through the community? Given the opportunity, how would they modify or extend the representations? This study cannot begin to answer these questions.

Some evidence that representational guidance plays a role in the classroom is provided by Toth et al. (2002). This classroom study involved larger groups, unrestricted use of representations, and greater time on task. The study compared two

forms of guidance for inquiry: Belvedere's Graph representations of evidential relations and assessment rubrics. The assessment rubrics were paper-based charts that included detailed criteria for progress in data collection, evaluation of information collected, quality of reports, and quality of peer presentations. The rubrics were provided to students at the outset of the study with explicit instructions concerning their use during the activity to guide inquiry. A 2×2 design crossed Graph (Belvedere) versus Text (Microsoft Word) conditions with Rubric versus No-rubric conditions, distributed across four ninth grade science classes working for about 2 weeks on each of three inquiry problems. For logistical reasons (the schools were located in Europe), data analysis was based primarily on artifacts produced by groups of students, namely their Belvedere graphs or Word documents and their final reports. (The lack of access to process data was one of the motivations for conducting this laboratory study.) Significant results were obtained on the epistemological categorization of information (H1, which was not significant in this study) and the number of inferences recorded in the form of evidential relations. The combination of graphing and rubrics resulted in a larger number of inferences recorded compared to all other conditions; the use of either graphing or rubrics alone did not result in a significantly higher performance compared to either of text groups. Interestingly, this interaction was primarily due to the Graph/Rubrics students having recorded significantly more inconsistency relations. Thus there appears to be a synergistic effect between effective representations and guidelines for their use, particularly with respect to attending to discrepant evidence. The rubrics encourage students to look for and record disconfirming as well as confirming information and Belvedere provides explicit representational devices for recording such information.

FUTURE WORK

There is of course a great deal of future work suggested by the study reported here, ranging from further analysis of existing data to new studies. The analyses presented here only assess the extent to which participants represented, discussed, and revisited items; they say little about the quality or depth of the collaborative knowledge-building activity. Subsequent analysis of our data will shift from comparison of group means to sociolinguistic methodologies (e.g., Interaction Analysis; Jordan & Henderson, 1995) to better understand how the different representations are appropriated as resources in support of collaborative discourse.

Follow-up studies within the experimental paradigm could investigate whether similar representational effects on collaboration are obtained in different tasks domains, with different learning or problem-solving objectives, and with other populations. At this writing we are concluding a small study to compare the co-present collaborations reported here with synchronous distance collaboration.

However, the most important follow-up may be undertaken by observing learning processes in schools or learning communities rather than laboratory experiments. Major limitations of the present research include the artificial nature of the task and setting and the limited temporal extent of participants' use of the representations. We have only established the presence of representational effects during initial use of representations on a limited task in a laboratory setting. Future work should be undertaken to investigate how learners' use of representations develops over time in an authentic inquiry setting and to explore whether the choice of representational notation also influences collaborative learning processes after extended use. Such investigations would better inform the development of instructional strategies for use of multirepresentational tools in realistic educational settings. The study is also limited to representations invented for the purposes of supporting basic generalized forms of evidence-based inquiry. Learners becoming enculturated into a community of practice encounter specialized representations that are integral to the practices of that community and carry with them conventional semantics (e.g., Hundhausen, 2002; Kozma & Russell, 1997). In such a situation, development of competency with these representations is an imperative, and social scaffolding may interact with or overcome effects due to representational affordances. This point underlines the need for study of how representational guidance might play out in the authentic inquiry of aspiring practitioners.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation under Grant 9873516.

Laura Girardeau ran the study sessions, transcribed and coded the videotapes, and gathered much of the data analyzed in this article. Michelene Chi provided valuable guidance and advice on the design and analysis of the experiment. Martha Crosby helped us with the statistical analysis. Bo Yang, Hongli Xiang, and Bin Ma prepared and maintained the experimental software. Arlene Weiner developed earlier versions of the materials on Mass Extinctions and Guam ALS-PD used in this study. Three anonymous reviewers suggested improvements to the presentation of this work. "Mahalo nui loa" to all of these individuals!

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

- Baker, M., & Lund, K. (1997). Promoting reflective interactions in a CSCL environment. *Journal of Computer Assisted Learning, 13*, 175-193.

- Bell, P. (1997). Using argument representations to make thinking visible for individuals and groups. *Proceedings of the 2nd International Conference on Computer Supported Collaborative Learning (CSCL'97)* (pp. 10–19). Toronto, Canada: University of Toronto.
- Brown, A. L., & Campione, J. C. (1994). Guided discovery in a community of learners. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and practice* (pp. 229–270). Cambridge, MA: MIT Press.
- Chi, M., Bassok, J., Lewis, M., Reimann, P., & Glaser, R. (1989). Learning from examples via self-explanations. In L. Resnick (Ed.), *Knowing, learning and instruction: Essays in honor of Robert Glaser* (pp. 251–282). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). Hyattsville, MD: American Psychological Association.
- Collins, A., & Ferguson, W. (1993). Epistemic forms and epistemic games: Structures and strategies to guide inquiry. *Educational Psychologist*, 28(1), 25–42.
- Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11, 671–684.
- Guzdial, M. (1997). Information ecology of collaborations in educational settings: Influence of tool. In *Proceedings of the 2nd International Conference on Computer Supported Collaborative Learning (CSCL'97)* (pp. 91–100). Toronto, Canada: University of Toronto.
- Guzdial, M., Hmelo, C., Hubscher, R., Nagel, K., Newstetter, W., Puntambekar, S., Shabo, A., Turns, J., & Kolodner, J. L. (1997). Integrating and guiding collaboration: Lessons learned in computer-supported collaborative learning research at Georgia Tech. In *Proceedings of the 2nd International Conference on Computer Supported Collaborative Learning (CSCL'97)* (pp. 91–100). Toronto, Canada: University of Toronto.
- Hoadley, C. M., Hsi, S., & Berman, B. P. (1995). The multimedia forum kiosk and SpeakEasy. In P. Zellweger (Ed.), *ACM Multimedia '95* (pp. 363–364). San Francisco: ACM.
- Hundhausen, C. D. (2002). Integrating algorithm visualization technology into an undergraduate algorithms course: Ethnographic studies of a social constructivist approach. *Computers & Education*, 39, 237–260.
- Jordan, B., & Henderson, A. (1995). Interaction analysis: Foundations and practice. *Journal of the Learning Sciences*, 4(1), 39–103.
- Koedinger, K. (1991). On the design of novel notations and actions to facilitate thinking and learning. In *Proceedings of the International Conference on the Learning Sciences* (pp. 266–273). Charlottesville, VA: Association for the Advancement of Computing in Education.
- Kotovsky, K., & Simon, H. A. (1990). What makes some problems really hard: Explorations in the problem space of difficulty. *Cognitive Psychology*, 22, 143–183.
- Kozma, R., & Russell, J. (1997). Multimedia and understanding: Expert and novice responses to different representations of chemical phenomena. *Journal of Research in Science Teaching*, 34, 949–968.
- Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, 11(1), 65–99.
- Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge, England: Cambridge University Press.
- Novak, J. (1990). Concept mapping: A useful tool for science education. *Journal of Research in Science Teaching*, 27, 937–49.
- Novick, L. R., & Hmelo, C. E. (1994). Transferring symbolic representations across nonisomorphic problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 1296–1321.
- Puntambekar, S., Nagel, K., Hübscher, R., Guzdial, M., & Kolodner, J. (1997). Intra-group and inter-group: An exploration of learning with complementary collaboration tools. In *Proceedings of the 2nd*

- International Conference on Computer Supported Collaborative Learning (CSCL'97)* (pp. 207–214). Toronto, Canada: University of Toronto.
- Roschelle, J. (1994, May). Designing for cognitive communication: Epistemic fidelity or mediating collaborative inquiry? *The Arachnet Electronic Journal of Virtual Culture*, 2(2). Retrieved from <http://www.kovacs.com/EJVC/ejvc.htm>
- Salomon, G. (Ed.). (1993). *Distributed cognitions: Psychological and educational considerations*. New York: Cambridge University Press.
- Scardamalia, M., & Bereiter, C. (1991). Higher levels of agency for children in knowledge building: A challenge for the design of new knowledge media. *The Journal of the Learning Sciences*, 1, 37–68.
- Schwarz, B., Neuman, Y., Gil, Y., & Ilya, M. (2003). Construction of collective and individual knowledge in argumentative activities: An experimental study. *The Journal of the Learning Sciences*, 12, 221–258.
- Slavin, R. E. (1990). *Cooperative learning: Theory, research, and practice*. Englewood Cliffs, NJ: Prentice Hall.
- Stein, B. S., & Bransford, J. D. (1979). Constraints on effective elaboration: Effects of precision and subject generation. *Journal of Verbal Learning and Verbal Behavior*, 18, 769–777.
- Stenning, K., & Oberlander, J. (1995). A cognitive theory of graphical and linguistic reasoning: Logic and implementation. *Cognitive Science*, 19(1), 97–140.
- Suthers, D. D. (1999a). Effects of alternate representations of evidential relations on collaborative learning discourse. In C. M. Hoadley & J. Roschelle (Eds.), *Proceedings of the Computer Support for Collaborative Learning (CSCL) 1999 Conference* (pp. 611–620). Palo Alto, CA: Stanford University. Available: <http://www.ciltkn.org/csc199/A74/A74.HTM>
- Suthers, D. D. (1999b). Representational support for collaborative inquiry. In *Proceedings of the 32nd Hawaii International Conference on the System Sciences (HICSS-32)* (CD-ROM). Maui, HI: Institute of Electrical and Electronics Engineers (IEEE). Available: <http://lilt.ics.hawaii.edu/lilt/papers/1999/Suthers-hicss99.pdf>
- Suthers, D. D. (2001). Towards a systematic study of representational guidance for collaborative learning discourse. *Journal of Universal Computer Science*, 7(3). Electronic publication: http://www.jucs.org/jucs_7_3/towards_a_systematic_study
- Suthers, D., Toth, E., & Weiner, A. (1997). An integrated approach to implementing collaborative inquiry in the classroom. In *Proceedings of the 2nd International Conference on Computer Supported Collaborative Learning (CSCL'97)* (pp. 272–279). Toronto, Canada: University of Toronto.
- Suthers, D., & Weiner, A. (1995). Groupware for developing critical discussion skills. In J. L. Schnase & E. L. Cunnius (Eds.), *Proceedings of the First International Conference on Computer Support for Collaborative Learning (CSCL'95)* (pp. 341–348). Mahwah, NJ: Lawrence Erlbaum Associates, Inc. Available: <http://lilt.ics.hawaii.edu/lilt/papers/1995/suthers-weiner-cscl95.pdf>
- Toth, E., Suthers, D., & Lesgold, A. (2002). Mapping to know: The effects of representational guidance and reflective assessment on scientific inquiry skills. *Science Education*, 86, 264–286.
- Webb, N., & Palincsar, A. (1996). Group processes in the classroom. In D. Berlmer & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 841–873). New York: Macmillan.
- Zhang, J. (1997). The nature of external representations in problem solving. *Cognitive Science*, 21, 179–217.