Instr Sci (2012) 40:19–46
DOI 10.1007/s11251-011-9168-3

Engaging young students in scientific investigations: prompting for meaningful reflection

Travis Wilson · Michelle Perry · Carolyn J. Anderson · Dean Grosshandler

Received: 30 September 2010 / Accepted: 11 May 2011 / Published online: 26 May 2011

Abstract This study examined the verbal prompts a tutor used to promote reflection and young students' responses to these prompts. Seven children (ages 8–12) participated in 260 min of one-on-one tutoring to learn scientific concepts related to gear movement; the tutor spontaneously provided these students with 763 prompts for reflection. Prompts reliably induced reflection: Students responded verbally 87% of the time. Turn-by-turn discourse analysis revealed seven distinct types of prompts and 11 distinct types of verbal responses. High-level prompts were strongly associated with high-level responses. A log-multiplicative association model with two dimensions (temporality and certainty) represented the relationships between prompt and response types; from this model, odds ratios estimated the strength of association between specific pairs of prompt and response types. Findings are discussed in terms of the effects that reflection may have on students' developing understanding of scientific concepts.

Keywords Reflection · Tutoring · Elementary science · Scientific problem-solving

Introduction

Learning scientific concepts has been notoriously difficult for many students. Although there is likely a host of reasons for this, we focus our concern on processes that might enable young students to overcome these difficulties. Among the myriad of possibilities for encouraging deep understanding, some of the most promising work has focused on encouraging students to reflect on their knowledge. The logic is that by facing both what they know and what they do not know, students can search for and integrate pieces of

T. Wilson (✉) · M. Perry · C. J. Anderson
Department of Educational Psychology, University of Illinois at Urbana-Champaign, 1310 S. Sixth St., Champaign, IL 61820-6990, USA
e-mail: mwalimu27@gmail.com

D. Grosshandler
University of Illinois at Chicago, 4138 SES, 848W. Taylor St., Chicago, IL 60607, USA

knowledge, leading to a much more sophisticated understanding than if they were not engaged in reflection.

In many ways, empirical evidence has largely demonstrated that reflection has many potential benefits: It has been linked to knowledge integration (Davis 2003), reducing the educational disadvantage of low-achieving students (White and Frederiksen 1998), and producing high conceptual gains among physics students (May and Etkina 2002). We get insight into how this works through Sandoval and Reiser's (2004) results. They found that planning and self-monitoring allowed students to identify and confront what they knew and what they did not know, thereby supporting students' representation and construction of scientific concepts. From this, we have evidence that reflection, especially when learning scientific concepts, may sustain quality scientific inquiry and improve learning outcomes.

The converse is also true: Learning devoid of reflection often lacks depth and conceptual clarity. When learners fail to engage in reflection, they tend to focus on products rather than on explanations (e.g., Schauble et al. 1995), are distracted by superficial features (Krajcik et al. 1998), and do not perform as well as students who engage in reflection (Chi et al. 1989; Recker and Pirolli 1995). Thus, a look at learning situations when reflection is or is not explicitly supported shows that when students receive support for reflection, they experience more positive learning outcomes (Songer and Linn 1991).

Indeed, reflection may be a critical aspect of the learning process (Reiser 2004). This may be especially true in scientific inquiry, relative to learning in other disciplines, given the prevailing emphasis on products rather than on principles in school-taught science (e.g., Krajcik et al. 1998). Furthermore, the cognitive load that accompanies the inherent difficulty in conducting complex investigations may lead to a lack of sufficient attention devoted to understanding the principles underlying the results (Loh et al. 2001). Given typical conditions for science learning, it is often difficult for students to form deep understanding of scientific principles and much more likely that they will form superficial, product-based understanding instead. As contemporary research indicates, reflection may play an integral role in reversing this typical pattern.

This interest in reflection is not new. For almost a century, researchers and educators have lauded the merits of encouraging students to use reflection in learning situations (Dewey 1916; Piaget 1967; Vygotsky 1962). The sociocultural perspective, in particular, suggests that reflection may have profound benefits for cognitive development, and that these benefits may be especially fruitful during middle childhood (7–12 years). Vygotsky (1962) claimed that an increase in reflective awareness and deliberate control is a cornerstone of cognitive development that occurs during middle childhood (and this claim is pervasive; e.g., see Piaget 1967). During this time period, attention and memory become less dependent on the immediate physical context, but rather more voluntary and increasingly dependent on the child's thoughts. Given this, we focus our attention on students in this age range.

But what, precisely, is meant by reflection? Like Davis (2003), we construe reflection to refer to both metacognitive and sense-making functions. That is, reflection can focus on the processes or quality of a student's thinking (metacognition), or on the particularities of subject matter being learned (sense-making). Meaningful reflection, therefore, calls attention to the ideas one takes for granted, how such thinking may be faulty or undeveloped, and how new experiences or novel concepts conflict or integrate with prior knowledge. In short, reflection promotes knowledge integration (Linn 1995).

Despite the reported successes of reflection in educational contexts, and the agreed-upon need for attention to reflection in science learning, research has shone insufficient light on the dynamic process of reflection itself (Davis 2003). Most studies that investigate reflection tend to conceive of reflection as an exercise to be done *after* a learning activity rather than *during* it. Schon (1992) distinguishes between these two types of reflection as reflection-*on*-action (reflecting on prior events) and reflection-*in*-action (reflection as learning transpires). Facility with reflection-*in*-action is necessary to monitor one's knowledge and sense-making as one's investigations gain complexity.

We argue that, in general, a better understanding of what prompts reflection *during* the learning situation and, more specifically, what sorts of conditions prompt what sorts of reflection, are issues that need to be addressed. Without a nuanced understanding of how to engage students successfully in reflection, and without an understanding of the range of conditions that might prompt both deeper and more superficial reflection, researchers lack traction either to advise teachers or to develop sophisticated models of the process of reflection.

Reflection-in-action: two critical dimensions

We posit that effective reflection has two distinct but synergistic dimensions in science learning. The first dimension concerns the instability of erroneous or incomplete knowledge: Reflection may induce uncertainty, which, in turn, allows the student to move beyond false or tenuous understandings to firmer conceptual ground (e.g., Perry et al. 1988). The second dimension concerns the organization of past, present, and future experiences (see Vygotsky 1978): Reflection allows one to compare prior knowledge with presently observed phenomena, and to test current hypotheses by making predictions about future events.

Uncertainty and knowledge in transition

Indexes of when someone is ripe for cognitive change typically incorporate ostensible measures of uncertainty, such as being vague and inarticulate (e.g., Graham and Perry 1993; Hosenfeld et al. 1997; Siegler and Jenkins 1989), making explicit metacognitive statements including "I don't think I understand" (e.g., Perry and Lewis 1999), or saying one thing, while providing discrepant information in gesture (e.g., Alibali 2005; Alibali and Goldin-Meadow 1993; Goldin-Meadow et al. 1993; Perry et al. 1988). Not surprisingly, these ostensible indexes are witnessed in response to a pointed question asking the student about a recent problem-solving attempt. In this way, asking students to reflect could either merely reveal or potentially provoke them to move to a state of uncertainty (Broaders et al. 2007; Cook et al. 2008). Moreover, the mere act of displaying uncertainty—especially in the presence of a tutor, teacher, or more sophisticated peer—has the added potential advantage of allowing the student to confront his or her own gaps in knowledge, thereby opening new possibilities to seek resolution or help in understanding (e.g., Sánchez et al. 2009). In this way, reflection may prompt students to reckon with new knowledge by exposing or addressing the student's uncertainty. The converse is also true: When students have a clear understanding about how and why something occurs, yet do not initially provide their reasoning, prompting can lead students to articulate their thoughts and facilitate a shared understanding between student and teacher (McNeill et al. 2006).

Knowing through time: past, present, and future

Any successful learning situation is at least in some ways dependent on what the student has previously learned. To produce successful learning outcomes, the student's knowledge state must be uncovered. Knowing what the student understands allows the instructor to extend beyond that already known knowledge. Uncovering naive understandings, misconceptions, or gaps in past knowledge allows the instructor to challenge these understandings or to fill in those gaps (Bouillion 2007). By assessing the student's current state of knowledge, the instructional enterprise can take root.

Elaborating partial understandings into more complete understandings also requires a student's full attention to what transpires during scientific investigations. Making detailed observations in the learning situation helps to identify previously overlooked phenomena that either substantiate or counter one's hypotheses. Likewise, monitoring one's own current thoughts enables efficient coordination between observing, thinking, and acting—investigation is more purposeful (Schraw et al. 2006). But these skills in observation and metacognition are difficult to acquire, especially among younger students (Brown and Campione 1990). Therefore, it is helpful to provide students with explicit encouragement to reflect on moment-to-moment events and cognitions to ground them more deeply in scientific exploration.

In teaching and learning in the domain of science, special attention must also be paid to the student's predictions of what will happen in the future. Hypothesis formation is a central tenet of the scientific enterprise. Thus, it is not enough to uncover what a student already knows, but also to guide the student to predict what will be. In general, these two aspects of learning and, more particularly, reconciling uncertainty with what is already known and what needs to be learned, likely arise as central components of most successful teaching-and-learning situations.

The present investigation

The present investigation attempts to address the complex micro-level processes of young students' reflection. In particular, we examined how a knowledgeable adult tutor induced novice students to use reflection to advance their scientific inquiry into the mechanics of gear movement. Various studies have shown that working with gears can be an especially rich source of insight into children's cognitive development (Perry and Lewis 1999; Chambers et al. 2008; Lehrer and Schauble 1998; Metz 1991). Lehrer and Schauble (1998, p. 4) give a particularly detailed rationale for this; after discussing the prominence of the role of reasoning about structure and function in influential science standards documents, they state:

> [G]ears provide an easy entry point for learning, yet also incorporate some relatively deep scientific (e.g., torque, mechanical advantage) and mathematical (e.g., ratio, proportion) principles…On the one hand, the operation of gears is directly inspectable and involves no hidden parts, and the transmission of motion is visible. Research on the development of causal reasoning suggests that even preschoolers are well equipped to understand the simple transmission of motion (Bullock et al. 1982). On the other hand, the literature on naive physics forewarns that developing a robust understanding of how motion is transmitted may require a considerable amount of inference, which in some cases may be governed by misapplied intuitive understandings (diSessa 1993).

The context of one-on-one tutoring seemed a natural fit for studying the processes of student reflection because it provides a context in which students may be asked to do a great deal of reflection, without which a deeper understanding of gears was not likely to be created (Lehrer and Schauble 1998). Engaging students in tutoring, although not necessarily aligned with actual scientific investigative practices and epistemology (Chinn and Malhotra 2002), nevertheless provides access to several important aspects of doing science: Students have opportunity to engage in close and relatively unguided observation, hypothesis formation, multiple trials, evidence-based reasoning, prediction, and (at least for older students) theory-building. Moreover, tutoring is highly effective. Studies have shown that the effect size of the advantage of tutoring over classroom learning has ranged from 0.4 to 2.3 standard deviations (Bloom 1984; Cohen et al. 1982; Mohan 1972). We believed that the processes of reflection would be magnified and more easily studied in tutoring situations as compared to other learning situations.

Choosing a knowledgeable adult tutor rather than a peer or a novice adult volunteer also was a strategic, though less obvious, choice. On one hand, as a meta-analysis of 52 tutoring studies has shown, a tutor's expertise is not significantly related to the impact of tutoring on learning (Cohen et al. 1982). Moreover, highly skilled tutors are relatively rare in most school systems. Perhaps for these reasons, studying naturalistic tutoring with non-expert tutors has received considerable attention (e.g. Graesser 1992; Graesser and Person 1994; Person et al. 1994). On the other hand, to learn about the *potential* that tutoring can have for the development of scientific knowledge, some expertise on the part of the tutor is necessary. Thus, we selected a tutor who was knowledgeable both in the mechanics of gear movement and how to prompt reflection as a pedagogical device (see also Chi et al. 2008; Putnam 1987).

In the present investigation, we explore five interrelated issues: (1) *How frequently* do prompts for reflection naturally occur in scientific, investigative tutoring situations? (2) What *types of prompts* does an expert tutor use to encourage student reflection and how do these prompt types differ substantively from each other? (3) How do *students respond* when provided with prompts for reflection? and (4) Is there a *relationship between the types of prompts received and the responses* students give? By illuminating these processes, our goal is to provide new understanding of how students can be prompted to engage in and reflect on learning about the mechanics of gear movement.

Method

Participants

Students

Seven children, aged 8–12 years, were recruited from the local Boys and Girls Club, an after-school program serving primarily low-income students. Of these seven children, six were boys (three African-American, three European-American) and one was a girl (African-American). Participants included all children whose presence at the program was fairly dependable, whose parents could be reached, and whose parent granted permission to participate. The students were not remarkable in their knowledge of science and math. All seven students had played with Legos before in some context—at a friend's house, at school, or in a store—but only two students owned a personal set of Legos. Four of the seven students previously had played with Lego gears, but none had used Lego gears in a formal learning situation at school or in a club.

The tutor

A male graduate student in mathematics, science, and technology education (fourth author) with extensive experience in classroom teaching and tutoring in Lego-Logo technology served as the tutor for all seven students. At the time of the tutoring, he had worked extensively with university-K12 STEM partnerships as a teacher, tutor, and consultant in science and engineering education. The tutor had considerable experience helping young students learn about gears and associated mathematical concepts through his work leading gears and mathematics workshops for K-8 schools, as well as 10 years of experience running an after-school engineering lab.

Procedure

Investigators videotaped 11 sessions with these seven students in one-on-one tutoring. The duration of the individual tutoring sessions varied from 7 to 48 min and the total time each child spent with the tutor ranged from 18.0 to 72.9 min ($M = 37.4$, SD = 19.2). Participants' scheduling issues were the reason for the variation in number of and duration of sessions. In the beginning of each lesson, the tutor explained that the purpose of the tutoring session was to explore gear movement and the interaction of various gears.

The materials in the learning situation were Lego gears and accompanying accessories. These materials were chosen because working with gears was both challenging and engaging for many young students. Students were given an assortment of gears (1–4 cm in diameter) that had 8, 24, or 40 teeth and short (8–10 cm long) plastic axles, which they could insert into the center of any gear. Students could turn a gear by touching it directly or by rotating the axle. Students were given a plastic housing unit into which they could insert the axles so that the gears could mesh easily at a fixed distance that could be selected by the student. The housing unit played an essential role because it allowed students to fit different combinations of gears together without difficulty. A student could then set into motion a pair of gears, or even a chain of three or more connected gears, by rotating just one gear. Figure 1 illustrates one such arrangement. Each gear was marked with a red dot on one of its teeth to enable participants to count rotations easily.

Investigating principles of gear movement

The tutor sought timely moments to guide each student towards learning three principles of gear mechanics: directionality of rotational motion, velocity ratio (VR), and mechanical advantage. In this section, we describe these principles in detail so that readers may have a common understanding of what the tutor hoped students would learn.

Typically, the first principle students encountered in the tutoring sessions was that interlocking gears cause each other to rotate in opposite directions. That is, if one gear spins in a clockwise direction, the other gear must spin in a counterclockwise direction, and vice versa. Although the directionality of gear motion may seem to readers to be a trivial principle, both young students and college students often misunderstand this phenomenon (Perry and Elder 1997; Perry and Lewis 1999).

The second principle students encountered was VR. VR applies to all simple machines and is the ratio of the distance moved by the point of application of the effort to the distance moved by the load. In the case of gears, VR also is an index that compares the rotational (or angular) velocity of two interlocking gears, where:

Fig. 1 Tutoring setting and example configuration of gears. *Note*. *Picture* at left illustrates the typical seating arrangement for the tutor and student; the gear set is located on the table closer to the student. Each student was given a set of 8-tooth, 24-tooth, and 40-tooth gears. The *picture* at right illustrates an example configuration of gears (an 8-tooth and a 24-tooth gear)

$$\mathrm{VR} = \frac{\text{Number of teeth on driver gear}}{\text{Number of teeth on follower gear}}$$

For example, if the driver gear has 48 teeth and the follower gear has 24 teeth, then the VR is $48/24 = 2.0$, where the follower gear rotates 2.0 times per rotation of the driver gear.

The third principle students encountered was mechanical advantage. Mechanical advantage is the factor by which a mechanism multiplies the force put into it to move a load. Generally, mechanical advantage is calculated as:

$$\mathrm{MA} = \frac{\text{Output force}}{\text{Input force}}$$

In the case of gears, mechanical advantage also is calculated as:

$$\mathrm{MA} = \frac{\text{Number of teeth on follower gear}}{\text{Number of teeth on driver gear}}$$

The following proportion summarizes these two relationships:

$$\mathrm{MA} = \frac{\text{Number of teeth on follower gear}}{\text{Number of teeth on driver gear}} = \frac{\text{Output force}}{\text{Input force}}$$

Thus, whenever the driver is connected to a relatively larger gear, force is multiplied; the follower rotates slowly compared to the driver but exerts a greater force. The reverse is true whenever the driver is connected to a relatively smaller gear. Students were not asked to formally calculate the mechanical advantage of a gear combination; rather, the tutor strived to develop in students an intuitive sense that a small gear, though little in stature, can yield considerable force when combined with a larger gear.

Although the tutor's goal was to have the students learn these three principles of gear mechanics, he did not strictly adhere to a pre-specified sequence of instruction. At times, the tutor guided students' activities. Other times, the tutor gave students latitude to explore their own ideas, encouraged them to ask questions, and frequently prompted students to reflect on the learning situation.

In our analyses, we focused on the verbal interactions between student and tutor insofar as these observations informed us about young students' processes of reflection. Thus, we coded two aspects of the tutoring session: (a) the tutor's prompts for student reflection and (b) students' verbal responses to prompts for reflection.

Coding

Both audio and video data were collected during each tutoring session via multiple microphones and cameras. Transcripts of the audio data were typed and served as the basis for coding dialogue between tutor and student.

Transcripts were marked for each turn. A turn of dialogue was defined as any utterance, phrase, question, or complex statement made by the tutor or student. Any response to a turn made by the other person was coded as a new turn. When a turn by either the tutor or student was followed by silence lasting 4 s or longer, a new turn was coded. The first and second authors used a subset of data (i.e., roughly 40 min) to develop the coding scheme.

Crafting the coding scheme was a hybrid of a priori and inductive processes. We knew in advance that we wanted to detect various types of prompts for reflection that were distinguishable from other types of tutoring discourse, such as simple fact-checking questions (e.g. "What is 7 times 11?") and didactic statements. We also expected there to be reflective prompts and responses of both higher and lower levels of rigor or quality. However, the types of prompts and responses that were eventually coded were not stipulated a priori, but emerged from our engagement with the transcripts in an inductive fashion.

After the coding scheme was fully developed, the first and second authors coded the entire corpus of data. The coding process had two basic steps. For the first step, we examined each tutor turn and decided whether or not it was a prompt for student reflection. We only coded reflective prompts and the responses to these prompts. For each prompt for reflection, we coded both the type (described in the next section; see Table 1) and level (high or low) of the prompt. For the second step, we identified the student turn that immediately followed the tutor prompt and classified the response's type (described later; see Table 2) and level (high or low).

There was one exception to the coding process: Whenever the tutor prompted a student for reflection but then interjected with another verbal statement before the student could respond, we classified the response as a "tutor interruption." Tutor interruptions occurred in only 50 out of the 763 response turns that followed reflective prompts (i.e., 6.6%).

Coding the tutor's prompts for reflection

In our corpus of tutor prompts, we found seven different types of tutor's prompts for reflection, which we list and describe in Table 1. We considered each prompt independently of the nature of the student's response to it. For example, if the tutor gave a prompt for a prediction, yet the student responded with a metacognitive comment, we coded the prompt as a request to make a prediction (and not a prompt for metacognition).

As a second level of coding, we classified each prompt as high- or low-level in complexity, according to the substance of the request. Low-level prompts made simple requests for descriptive information about a physical event, to recall a prior event, or to restate something that was previously said (i.e., akin to remembering and understanding, per Bloom's taxonomy; Bloom 1956). High-level prompts asked the student to interpret, explain, predict, generalize, synthesize, or evaluate information (i.e., akin to applying, analyzing, evaluating, or creating, per Bloom's taxonomy; Bloom 1956). At this level of coding, too, we considered each prompt independently of the nature of the student's

Table 1 Description of coding for types of prompts for reflection

Code	Description
Reflection on action	Requests student to consider gear action or student's manipulation of the gears. E.g. "What just happened? What did the gear do?"
Metacognitive	Asks student to reflect on his/her own thinking. E.g. "Were you aware of what you were doing?"
Metalinguistic	Prompts student to analyze his/her verbal statements. E.g. "What do you mean when you say…?"
Real world connection	Leads student to make a connection between the learning situation and real world events or materials. E.g. "Have you seen something like this before at home or at school?"
Explanations	Checks for understanding or requests a synthesis of concepts. E.g., "How does the size difference between gears reflect their relative strength?"
Apply ideas	Encourages the student to apply or test an idea by manipulating the gears. E.g. "How can you test that idea with the gears in front of you?"
Prediction	Requests the student to make a prediction about what might happen under given conditions. E.g. "What will happen if…"

Table 2 Description of coding for students' responses to prompts

Code	Description
Silence	Any time the student fails to make a verbal response, except when he or she clearly makes a statement through gesture (e.g. nodding)
Question	Student asks a question of any kind, however simple or complex
Prediction	A prediction of what would happen if the student were to apply a particular action to the gears
Hypothesis	Any conjecture about the principles believed to underlie the gears' observed behavior
Yes/no	A "Yes" or "No" statement or any equivalent head movements
Impasse	An expression of uncertainty or confusion. E.g. "I don't know"
Real world	A reference made to real world objects or events outside of the learning situation. E.g. "It looks like an engine belt"
Making plans	Statements of what to do next with the gears or how to do it. E.g. "Let's try doing it the opposite way"
Simple description	Simple accounts of what one did, experienced, or saw happen. E.g. "This gear went clockwise really fast"
Complex description	Provides layers of description or genuine insight about how one interprets an experience or observation. E.g. "The second pair of gears were more like each other in speed than the first pair of gears"
Emotional response	Overt signs of affective responses. E.g. "Oh my! Wow!"
Make meaning	Makes connections to lessons learned previously or otherwise attempts to make meaning. E.g. "I know that the small one is stronger than the big one, but I am wondering if size matters of if the number of teeth matters"
Tutor speaks	When the tutor interjects before the student has a chance to respond to the previous prompt

response to it. Each prompt type could have high- and low-level forms. For example, a meta-linguistic prompt would be classified as low-level (or high-level) if the student were simply asked to paraphrase (or evaluate) what s/he previously said.

Coding students' responses to prompts for reflection

To determine students' responses to prompts for reflection, we coded the single turn that occurred immediately after each prompt. We reliably found 13 different response types (including the infrequently occurring tutor interruptions). We list and describe these in Table 2. As a second level of coding, we classified each response as high- or low-level in complexity. Low-level responses included yes/no statements, expressions of uncertainty and confusion, or simple descriptions of events or ideas (i.e., akin to remembering and understanding, per Bloom's taxonomy; Bloom 1956). In contrast, high-level responses indicated that the student had compared or contrasted pieces of knowledge, provided genuine insight, or articulated a meaningful conjecture (i.e., akin to applying, analyzing, evaluating, or creating, per Bloom's taxonomy; Bloom 1956).

To help the reader better grasp the nature of the interactions between the tutor and the student, we provide a sample excerpt from a transcript of one tutoring session. In this scenario, the tutor furnished the child with two gears interlocked, a 24-tooth gear and a 40-tooth gear. He wanted the child to witness the phenomenon of VR, the ratio of the linear speeds of the teeth on the respective gears. To begin, the tutor simply asked the child to spin the gears and to pay attention to their motion.

T: So, what do you notice when you spin them? (P: ACTION—LOW)
C: [*First the child spins the 40-tooth gear, causing the 24-tooth gear to move as well.*] They go round, and they kind of move real fast. [*Now he spins the 24-tooth gear, setting in motion the 40-tooth gear.*] (R: *Simple description—low*)
T: Do you notice? You are doing something differently. 'Cause you are not, they don't look the same way when you spin the 24-tooth gear. (P: ACTION—LOW)
C: It goes a bit faster than when I use the little one. 'Cause the little one doesn't have enough teeth for to help them spin around… (R: *Complex description—high*)
T: Say that one more time? The little one… (P: METALINGUISTIC—LOW)
C: The little one doesn't have enough, enough teeth to help them spin around a lot faster as to the 40-tooth…or the 40-tooth gear. (R: *Make meaning—high*)
T: So, what would happen if you put the 40-tooth gear and the 8-tooth [together]?
C: [*The student reaches for the 40-tooth gear.*] 40-tooth…
T: Yep. Before you do it, tell me what you think would happen. (P: PREDICT—HIGH)
C: It will probably [be] the same except a lot, a bit faster because of less weight of pushing on, holding, um, doing it also. (R: *Predict—high*)
T: Okay. Let's try it. (P: APPLY—LOW)

This excerpt is typical of the entire corpus in that the student received multiple prompts for reflection in a relatively brief amount of time (57 s in the transcript excerpt provided here). This scenario also provides contextualized examples of several prompt types: an ACTION-LOW prompt (e.g. "So what did you notice when you spin them?"); a METALINGUISTIC-LOW prompt ("Say that one more time? The little one…"); and a PREDICTION-HIGH prompt ("Before you do it, tell me what you think would happen.").

Inter-rater reliability

After completing the coding for prompts and responses for the entire corpus, a third person (not an author) coded roughly 25% of the total minutes, across all participants, to establish inter-rater reliability. As measured by kappa values, inter-rater reliability was deemed good for all aspects of coding. For prompt type, kappa = 0.75; for prompt level, kappa = 0.77; for response type, $\kappa = 0.77$; and for response level, $\kappa = 0.73$.

Results

Our initial goal here is to describe what happened between tutor and student, in terms of getting students to reflect on physical principles of gear movement. We first provide descriptive statistics on the prompts and responses. Next, we provide initial sequential analyses to examine the relations between prompts and responses. After we provide descriptive and sequential statistics on this issue, we turn to models for categorical data analysis to explore more deeply how these tutoring sessions might impact student reflection; given our small sample, we use these models for descriptive rather than inferential purposes. In addition, we hope to use these analyses to generate hypotheses about the relationship between tutor prompts and student attempts to reflect on the learning situation.

Descriptive statistics

In this section, we begin by describing the tutor's prompts. We first report the frequency of different types of tutor prompts. To explore the general variability of the prompts experienced by each student, we provide an index of heterogeneity of the tutor's prompts. We then look at the relative frequency of low- versus high-level prompts.

Next, we turn to the students' responses. We begin by reporting the frequencies of different types of responses. As with the tutor's prompts, we also examine how diverse these responses are. In particular, we examine response diversity in terms of response type, high- versus low-level, and engagement with the gears.

Frequency of tutor prompts for reflection

How frequently did the tutor provide prompts for the students to reflect? The tutor had 1,682 turns in the tutoring dialogue. Of these, we found 763 prompts for reflection (45.4% of all tutor turns) in 260 min of tutoring, or an average of 2.9 prompts for reflection per minute (SD = 0.77). We show the number of prompts for reflection provided to each child in Table 3.

Diversity of prompt types

If prompts for reflection are a tool that assists scientific inquiry, then a diversity of prompts may be crucial to elicit a wide range of cognitive processes integral to this endeavor. How many of each type of prompt did students receive? In Table 4 we present the frequencies of each prompt type provided to each student. As the last row in the table illustrates, there was considerable diversity in the types of prompts students received: No single prompt type accounted for more than 24% of all prompts for reflection. However, prompts for reflection on actions and for application together comprised nearly 46% of all prompts, showing that the tutor frequently encouraged students to attend directly to the physical details of gear movement and gear manipulation as a pedagogical strategy. The tutor also regularly promoted self-regulation by encouraging students to monitor their own thinking, as metacognitive and metalinguistic prompts together accounted for nearly 17% of all cases.

Table 3 Frequency of prompts for reflection per student

Student	Total length of corpus (min)	Number of prompts for reflection	Prompts per minute
Frank	19.0	64	3.4
Enrique	46.0	109	2.4
Damion	72.9	231	3.2
Chris	43.9	126	2.9
Ricardo	35.5	144	4.1
Charles	24.7	56	2.3
Sophia	18.0	33	1.8
Total	259.9	763	2.9

Table 4 Frequencies of types of prompts for reflection per student

Student	Action	Meta-cognitive	Meta-linguistic	Real world	Explain	Apply	Predict	Total	Prompt diversity
Frank	14	5	1	25	9	6	4	64	0.761
Enrique	31	16	10	18	9	19	6	109	0.822
Damion	37	37	14	21	36	45	41	231	0.843
Chris	38	13	8	8	17	35	7	126	0.792
Ricardo	30	8	7	16	26	37	20	144	0.821
Charles	20	7	0	5	6	12	6	56	0.780
Sophia	8	1	1	0	2	19	2	33	0.601
Total	178	87	41	93	104	174	86	763	0.774
Percent	23.3	11.4	5.4	12.2	13.6	22.8	11.3		

How might we characterize the distributions of prompts students receive? First, almost every student received every type of prompt, although each student received them in varying amounts (Pearson's $\chi^2(36) = 139.9, p < 0.01$). Next, we thought it would be illuminating to explore the general variability of prompts experienced by each student. Simpson's (1949) Diversity Index provides us with a means to quantify prompt heterogeneity for each student. Simpson's Index takes into account both richness (the presence of different types of prompts) and evenness (the relative abundance of each prompt type). The index ranges from 0 to 1, where increasing values denote greater heterogeneity in prompt types.

"Prompt diversity" in the final column of Table 4 presents Simpson's Index for each student. In our corpus, prompt diversity was quite high and this was similar across all students (but was less true with the last student listed in Table 4). Therefore, although the precise distribution of types of prompts differed among students, the students nonetheless experienced similarly rich discourse with the tutor, as marked by a wide diversity of types of prompts for reflection. From an analytical view, these similarities across tutoring sessions provide some evidence that supports the aggregation of student data in subsequent statistical analyses.

Low- and high-level prompts

Next, we examined the relative frequency of low- versus high-level prompts. The relative preponderance of low- versus high-level prompts stands as an index of the depth of

Table 5 Percent of low-level and high-level prompts for reflection per student

Student	% Low-level prompts	% High-level prompts
Frank	68.8	31.3
Enrique	71.6	28.4
Damion	65.4	34.6
Chris	76.2	23.8
Ricardo	70.1	29.9
Charles	85.7	14.3
Sophia	87.9	12.1
Total	71.7	28.3

scientific thinking the tutor asked of his students. How prevalent was each prompt level? In Table 5, we present the percentages of low- and high-level prompts for reflection received by each student. On average, 71.7% of the tutor's prompts were lower level and 28.3% were higher level. Each student received significantly more low-level prompts than high-level prompts (all Pearson's $\chi^2(1) > 9.6$, $p < 0.01$). The relative preponderance of low- versus high-level prompts notwithstanding, the tutor regularly challenged each child with a substantial number of high-level prompts for reflection, which potentially elevated his or her thinking to higher levels of rigor or abstraction. The high-level prompts also afforded the possibility of making the student's understandings and misconceptions explicit to both tutor and student.

Responses to prompts for reflection and diversity of response types

How did students actually respond to the prompts for reflection? Overall, students provided verbal responses to 86.6% of the tutor's prompts to reflect. At the very least, then, prompts for reflection were an excellent way to promote student engagement in the learning situation. In Table 6, we present a cross-tabulation of prompt types with response types that summarizes how students responded to different types of prompts. In 50 cases, the tutor interjected before the student could respond to a prompt. In two other cases, the student responded to a prompt with an emotional utterance (e.g. "Oh my!") that lacked substantive information. We removed these 52 cases from the summary table, leaving a total of 711 responses for our analyses.

As can be seen in the bottom row of Table 6, there was an abundance of cases of each type of response in our corpus, which stands as further evidence that prompts for reflection can lead to diversity in cognitive engagement with the learning situation. Similar to our description of prompt diversity in Table 4, the final column in Table 6 indexes the diversity of responses for each prompt type given. The average response diversity ($d = 0.77$) is quite high, although values are generally lower for prompts that request students to make connections to the REAL WORLD ($d = 0.62$) and prompts that request students to make PREDICTIONS ($d = 0.63$). Prompts for EXPLANATIONS ($d = 0.85$), METACOGNITION ($d = 0.83$) and to APPLY ideas ($d = 0.83$) yielded the greatest response diversity.

Low- and high-level responses

Overall, students responded to prompts much more frequently with low-level responses ($n = 484$) than with high-level responses ($n = 227$). But was the level of students'

Table 6 Cross-tabulation of prompt types versus response types for all students

Prompt type	Response type											Total	Diversity index
	Yes/no	Impasse	Real world	Make plans	Describe (simple)	Describe (complex)	Make meaning	Silence	Ask question	Make prediction	Hypo thesis		
Action	33	6	1	3	47	42	5	12	5	0	4	158	0.787
Meta- cognitive	9	6	2	2	10	8	25	14	6	0	1	83	0.834
Meta- linguistic	3	1	2	1	7	12	6	0	2	0	3	37	0.812
Real world	26	2	49	1	1	2	1	2	5	0	2	91	0.623
Explain	24	9	2	1	19	7	19	9	6	2	3	101	0.847
Apply	10	9	1	8	23	11	7	51	28	6	4	158	0.825
Predict	2	4	0	1	5	4	3	7	6	49	2	83	0.628
Total	107	37	57	17	112	86	66	95	58	57	19	711	0.765
Percent	15.0	5.2	8.0	2.4	15.8	12.1	9.3	13.4	8.2	8.0	2.7		

Table 7 Frequencies of high- and low-level responses by high- and low-level prompts for all students

Prompt level	Response level		
	Low	High	Total
Low	401	107	508
High	83	120	203
Total	484	227	711

responses contingent on the level of difficulty of the tutor's prompts? In Table 7, we present the frequency of low- and high-level prompts that led to either low- or high-level responses. Indeed, students responded to low-level prompts with more low-level (79%) than high-level (21%) responses and responded to high-level prompts with more high-level (59%) than low-level responses (41%).

To determine whether the relationship between prompt level and response level was similar across students or whether students responded uniquely, we conducted the Breslow-Day test for homogeneous association. The test statistic was not significant ($\chi^2 = 6.34$, $p = 0.39$, $n = 711$). From this, we inferred that odds ratios between prompt level and response level are equal over students, indicating relatively similar behavior across all seven students. We then conducted the Cochran–Mantel–Haenszel test to determine whether there was a relationship between prompt and response, across students. The test statistic was significant (82.74, $p < 0.01$) and estimated that the common odds ratio equals 5.12 (95% CI: 3.54–7.39). This means that the odds that a student will produce a high- versus low-level response when given a high-level prompt is 5.12 times the corresponding odds when the student is given a low-level prompt. These results, particular to elementary-school-aged students learning specific principles of gear mechanics, are consistent with other studies that have shown that higher-order questioning predictably facilitates higher levels of cognition (Ozgungor and Guthrie 2004) and performance (Rubie-Davies 2007).

Log-linear and log-multiplicative analyses[1]

We used log-linear models and related log-multiplicative association models to represent the relationship between prompt and response types. Based on our questions, we included in the model not only tutor prompts and student responses, but also individual students as a discrete variable. We did this because students differed in terms of the number of prompt types they received and the number of response types they made, so they might also have differed in the relationship they exhibited between prompt and response type. Log-linear models were fit to the three-way cross-classification of seven prompt types by eleven response types by seven students.

Zeros in table cells (in the present case, when a particular student either does not receive a particular type of prompt or does not present a particular type of response) can create problems for estimating interaction terms and assessing model goodness-of-fit. We found that we could not estimate two (out of the 49) student-by-prompt interaction terms and 10 (out of the 77) student-by-response interaction terms. To deal with this estimation problem, we used methodology for incomplete tables (Agresti 2002). To assess the goodness-of-fit, we relied on studying residuals, which should be approximately standard normal if a model

[1] SAS and LEM code for all log-linear and log-multiplicative models will be made available online at http://faculty.ed.uiuc.edu/cja/homepage.

provides a good fit to the data. All models were fit to the data using the LEM program (Vermunt 1997).

Homogeneous log-linear model

We found that the homogeneous log-linear model, which contains two-way interactions between each pair of variables, provided a good representation of the data. There were only two large standardized residuals (i.e., 5.24 and 3.30), which is reasonable given that there are 539 cells in the table. The fact that the homogeneous model gives a good representation of the data itself is noteworthy: It suggests that the relationships between prompt types and response types remained relatively stable across students. In other words, there were no three-way interactions (prompt x response x student) within the homogeneous log-linear model.

Final model: a log-multiplicative model

Because the relationship between prompt and response is of major interest and there are 77 interaction terms to represent this effect, we used a log-multiplicative association model. By doing so, we replaced the unstructured interaction terms in the log-linear model with the product of an association parameter and scale values for the prompts and responses. The product terms place restrictions on the unstructured interaction terms (see Becker 1989). For readers who are familiar with these models, we note that these models are extensions of the multidimensional row-column association model for two-way tables (Goodman 1981; see also Agresti 2002) to three-way tables. The log-multiplicative association models we used are special cases of the homogeneous log-linear model. The methodology required in the log-linear modeling to deal with sampling zeros was not required in the log-multiplicative association models.

We tried to fit log-multiplicative association models to the data with one and two dimensions, respectively, for the prompt–response interaction, with the goal of determining whether one- or two-dimensional models fit the data better. A one-dimensional model was not sufficient to fit the data; however, a two-dimensional model with only two large standardized residuals (i.e., 4.04 and 5.34) did suffice.

For the readers who want specifics of the two-dimensional log-multiplicative model, we explain the details in this paragraph. This model represents the interaction between prompt and response by $\varphi_1\mu_{1i}\nu_{1j} + \varphi_2\mu_{2i}\nu_{2j}$, where μ_{1i} and μ_{2i} are the scale values on dimensions one and two for prompt i; ν_{1j} and ν_{2j} are scale values on dimensions one and two for response j; and φ_1 and φ_2 are the association parameters for the two dimensions, respectively. The scale values and association parameters, as well as all other terms in the model, are estimated from the data. The scale values represent the structure or nature of the relationship between the prompts and response types, and the association parameters represent the strength of the relationship. In log-multiplicative association (and log-linear) models, the relationships between variables are defined in terms of odds ratios. The interaction model parameters give us estimates of these odds ratios. In particular, the odds ratio of response type j versus j' for prompt types i versus i' equals $\exp[\varphi_1(\mu_{1i} - \mu_{1i'})(\nu_{1j} - \nu_{1j'}) + \varphi_2(\mu_{2i} - \mu_{2i'})(\nu_{2j} - \nu_{2j'})]$.

We chose to represent the model results graphically, which we present in Fig. 2. Prompts are written in capital, non-italicized letters. Responses are written in lower-case, italicized letters both in the figure and in the text. Before moving to the interpretation of the

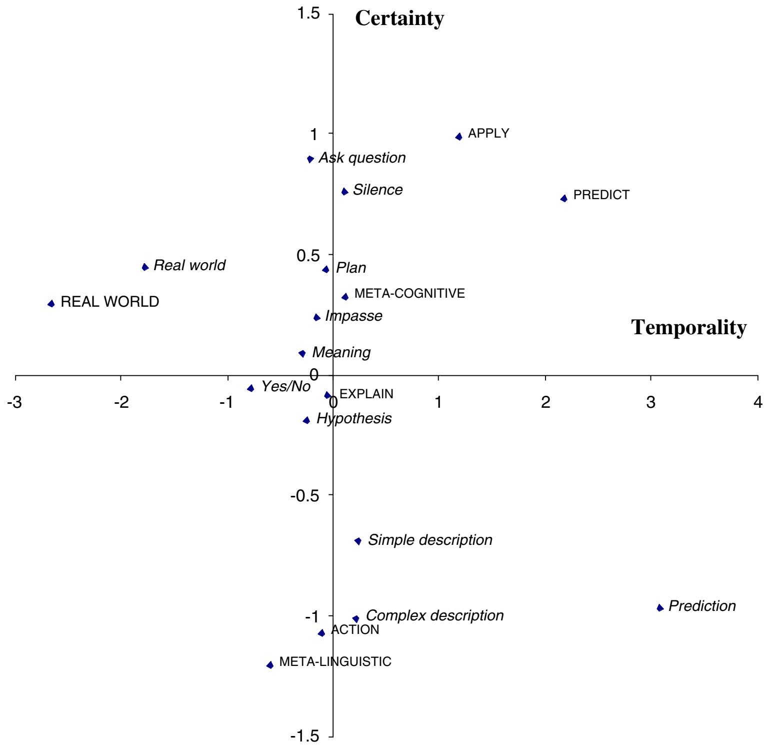

Fig. 2 Graphical representation of the log-multiplicative model of associations between prompt types and response types. *Note*. Prompts are written in *capital*, *non-italicized letters*. Responses are written in *lower-case*, *italicized letters*. The first dimension, depicted along the horizontal axis, represents time and moves from the past (far left), through the present (near the *center*), toward the future (far *right*). The second dimension of association, depicted along the vertical axis, represents level of certainty and moves from relative certainty (near the *bottom*) toward relative uncertainty (near the *top*)

figure, we first provide a technical explanation of how we produced Fig. 2. In this figure, we plot the estimated scale values for the prompts and responses weighted by $\sqrt{\varphi_1}$ and $\sqrt{\varphi_2}$. We plot Dimension 1 values, $\sqrt{\varphi_1}\mu_{1i}$ and $\sqrt{\varphi_1}v_{1j}$, on the horizontal axis and the Dimension 2 values, $\sqrt{\varphi_2}\mu_{2i}$ and $\sqrt{\varphi_{2j}}v_{2j}$,, on the vertical axis.

When interpreting the figure, note that the absolute value of the distances between points is not meaningful, but the relative distances are. If the points for two prompt (or response) types are relatively close to each other, then the odds ratios that involve these two prompts (or responses) are similar in value. The greater the difference between the scale values for two prompts or two response types, the further the odds ratio is from one, which indicates a strong association. For example, the points for prompts REAL WORLD and PREDICT are far apart in the figure, thus any odds ratios that involve these two prompts will be relatively large.

Interpreting the log-multiplicative model

To aid interpretation of the log-multiplicative model and the relationships between prompt and response types, we ascribed meaning to each of the two dimensions shown in Fig. 2;

this was done in a post-hoc manner after inspecting the general pattern of results in the graph. The first dimension, depicted along the horizontal axis, represents time; it moves from the past (far left), through the present (near the center), toward the future (far right). For example, REAL WORLD prompts and *Real World* responses refer to events and ideas that occurred *before* the tutoring session and are located along the left side of the first dimension. In contrast, PREDICTION prompts and *Prediction* responses, which refer to what might happen in the *future*, are located horizontally farther to the right. Other prompts and responses that pertain to events that just occurred within the learning situation are located horizontally near the center.

The second dimension, depicted along the vertical axis in Fig. 2, represents level of certainty. Three types of responses demonstrate substantial uncertainty: *Impasses* (expressing confusion or uncertainty), asking *Questions*, and *Silence*; these are found on the positive end of the vertical axis. Together, these three types of responses accounted for 27% of the total responses. These abbreviated responses reflect uncertainty on the part of the student both explicitly and implicitly; *Impasses* and *Questions*, by their very nature, are explicit expressions of uncertainty, while *Silence* is an implicit expression of uncertainty (e.g., Perry and Lewis 1999; Hosenfeld et al. 1997). Located at the opposite end of the second dimension (i.e., the negative end of the vertical axis) are responses reflecting greater certainty. *Simple* and *Complex Descriptions* are located near the bottom of the vertical axis. Such responses indicate that students are more self-assured, as when they describe events that have just occurred or observations they have just made.

The relative distances between prompts and responses have meaning in two distinct ways. First, the closer together two responses (or prompts) are, the more similar they were with respect to the two dimensions. For example, *Simple* and *Complex Descriptions* were more similar to each other than were *Real World* and *Prediction* responses.

Second, the closer together a prompt type is to a response type, the more likely the prompt type elicited that particular response type. Thus, a visual inspection of the graph reveals, for example, that REAL WORLD prompts were much more likely to yield *Real World* responses than any other type of response. Similarly, prompts to reflect on ACTIONS were more likely to yield *Complex* or *Simple Descriptions* than any other type of response (and, to take it one step further, *Complex Descriptions* slightly more so than *Simple Descriptions*).

The cluster of responses near the center of the graph are particularly noteworthy because they all refer to the present learning situation and seem to be marked by moderate to high levels of uncertainty on the part of the student. Prompts for EXPLANATIONS and METACOGNITIVE reflections are the prompt types nearest to these responses. It is reasonable to infer that these prompt types were particularly effective in leading students toward "shakier ground" where uncertainty and cognitive conflict may disrupt their equilibrium. As another benefit, prompts for EXPLANATIONS and METACOGNITION elicited a variety of types of responses.

Strength of association between particular prompts and responses: using odds ratios

The associations between some prompt and response types deserve closer attention. We argue that associations based on model parameters are more realistic than odds ratios based on raw counts, which ignore variability across students. Moreover, since the standardized residuals are relatively small, the model provides a good representation of the structure in the data. Thus, we calculated the odds ratios among prompt–response associations by using their corresponding values for $\sqrt{\varphi_1}\mu_{1i}$ (or $\sqrt{\varphi_1}\nu_{1j}$) and $\sqrt{\varphi_2}\mu_{2i}$ (or $\sqrt{\varphi_{2j}}\nu_{2j}$). We use odds ratios not as inferential statistics to generalize to a larger population, but as a means to succinctly describe meaningful associations in the sample data. To provide the reader with an intuitive sense of what the odds ratios mean, we first present two examples. We then describe a specific set of odds ratios that are particularly meaningful given our questions and the nature of our data.

In general, in computing an odds ratio, we consider together two prompts and two responses. One prompt and one response pair is of interest; that is, we want to know how strongly a particular prompt type and a particular response type were associated with each other. The other prompt and other response provide a sensible basis of comparison. For example, we may be interested in how to initiate student engagement and develop a verbal rapport between tutor and student; in other words, we want to know how to get students to become familiar with the learning situation—and not to become confused or overwhelmed. In this case, *Description* would be the response of interest because getting a student to provide a simple description gets the student engaged with the task in a relatively non-threatening way and *Impasse* would be an appropriate comparison because this would indicate that the student would be overwhelmed or confused. To select the prompts that could lead to this response, we may have an intuitive sense that both ACTION prompts and APPLY prompts often lead to *Simple Description* responses, but we suspect that APPLY prompts also lead to *Impasses*. Thus, ACTION prompts are the prompts of interest and APPLY prompts is the comparison. We calculate according to the following formula:

$\exp[\varphi_1(\mu_{1i} - \mu_{1i'})(v_{1j} - v_{1j'}) + \varphi_2(\mu_{2i} - \mu_{2i'})(v_{2j} - v_{2j'})]$, which yields an odds ratio of 3.99.

In this particular instance, we interpret the odds ratio to mean that the odds of a student producing a *Simple description* response versus an *Impasse* after the tutor gave an ACTION prompt were 3.99 times the corresponding odds as when he gave an APPLY prompt. Returning to the original question, we might further infer from this relatively large odds ratio that ACTION prompts were more conducive than APPLY prompts to easing the student into the learning situation and establishing a rapport between student and tutor than were APPLY prompts. There are no generally accepted levels for interpreting odds ratios, except that the farther the odds ratio from 1.0, the stronger the association between a given prompt type and a given response type.

The final analyses we report below elaborate on prompts and responses that featured important domains of effective science tutoring: (a) building scientific knowledge, (b) promoting rich discourse, and (c) revealing knowledge in transition. In doing so, we describe key prompts and responses, and the strength of their associations across all tutoring sessions (i.e., model-based odds ratios, summarized in Table 8).

Building scientific knowledge

We were interested in documenting specific types of scientific thinking and the types of prompts most closely associated with them. Focusing on detailed observations, connecting scientific concepts to real-world phenomena, and making informed scientific predictions are three ways that students can build and elaborate their scientific knowledge. The question then became: Which types of prompts best elicited these types of engagement with scientific thinking?

To begin, we asked what prompt would lead to a description of the actions and characteristics of the gears, which likely is necessary to develop a rich understanding of gear movement. We chose as our most likely candidate the tutor's ACTION prompts (e.g.,

Table 8 Summary of odds ratios of associations between key prompts and responses

	Compared prompts	Compared responses	Odds ratio
Building scientific knowledge			
Detailed observations	Action versus explain	Simple desc. versus impasse	2.4
Real world connections	Real world versus explain	Real world versus impasse	73.5
Scientific predictions	Prediction versus explain	Prediction versus impasse	534.3
Rich discourse			
Deeper meaning	Explain versus action	Make meaning versus simple desc.	2.1
	Metacognitive versus action	Make meaning versus simple desc.	2.6
Complex descriptions	Action versus apply	Complex desc. versus simple desc.	2.0
	Action versus metacognition	Complex desc. versus simple desc.	1.6
Knowledge in transition			
Impasses	Apply versus action	Question versus simple desc.	14.3
Silence	Apply versus action	Silence versus simple desc.	16.4
Summary of odds ratios of associations between key prompts and responses			
Questioning	Apply versus action	Impasses versus simple desc.	4.0

Note. Readers should interpret the odds ratios as descriptive, not inferential, statistics

"what did that gear do?"). For comparisons, we also selected requests for explanations (EXPLAIN prompts), as these prompts could easily lead students to describe the actions of the gears. The odds of eliciting a *Simple Description* response versus an *Impasse* response when the tutor gave an ACTION prompt were 2.4 times the corresponding odds when he gave a prompt to EXPLAIN. This odds ratio is substantially larger than 1.0, suggesting a strong association between ACTION prompts and *Simple Description* responses. (We expect that this ratio would have been even larger had we not disaggregated *Description* responses into *Simple* and *Complex Descriptions*.) This result indicates that ACTION prompts were an effective way to have students pay attention to the details of gear movement and articulate their thoughts.

Next, we asked what prompt type might lead students to draw a comparison from the experimental situation to their real-world experiences, as an indication that the students were making personal sense of the scientific problem. Obviously, asking students for a real-world connection should do this most effectively. When considering comparisons, we again chose explanation prompts, primarily for the same reason: A generic request for scientific explanation may lead the student to think about how the experimental situation connects to other phenomena he or she has experienced. The odds of yielding a *Real World* response versus an *Impasse* response when the tutor gave a REAL WORLD prompt were 73.5 times the corresponding odds when he gave a prompt to EXPLAIN. This high odds ratio shows how effective REAL WORLD prompts were at getting students to connect their ideas to prior life experiences and their intuitive senses of how the physical world operates: The student was much more likely to make real world connections when explicitly requested by the tutor to do so as compared to when the tutor simply asked for an explanation more broadly.

Finally, what kinds of prompts lead students to provide predictions? Of course, we expected or at least hoped that asking the student to predict would actually lead to predictions. But requesting predictions in actuality may result in impasses, or silence, or something else altogether. Moreover, it is possible that a more generic prompt, designed to engage the student in scientific thinking, may also lead to predictions, and perhaps more reliably so than specific requests for predictions. With this in mind, we chose requesting EXPLANATIONS as the comparison prompt and *Impasses* as the comparison response for computing odds ratios in associating PREDICTION prompts with *Prediction* responses. The odds of yielding a *Prediction* response versus an *Impasse* when the tutor gave a PREDICTION prompt were 534.3 times the corresponding odds when he gave a prompt to EXPLAIN. This extremely high odds ratio shows how PREDICTION prompts were very strongly associated with *Prediction* responses. We may infer that PREDICTION prompts were highly effective at getting students to attempt to project what they had learned in the tutoring session onto future, hypothetical scenarios.

Promoting rich discourse

We were also interested in documenting which prompts were most effective in leading students to engage in rich scientific discourse, as evidenced by students' attempts to *Make Meaning*. We wanted to contrast these substantive responses with a relatively thin response, but potentially related response, so we chose *Simple Descriptions* as the comparison. For the baseline comparison, we chose ACTION prompts because this was the most preponderant prompt type used by the tutor. To select the prompts that we wanted to investigate as being associated with *Make Meaning* responses, we chose both EXPLAIN and METACOGNITIVE prompts because these prompts by nature request students to think rigorously and to monitor their thoughts. The odds of eliciting a *Make Meaning* response versus a *Simple Description* when the tutor gave an EXPLAIN prompt were 2.1 times the corresponding odds when he gave a prompt to reflect on an ACTION. The odds of eliciting a *Make Meaning* response versus a *Simple Description* when the tutor gave a METACOGNITIVE prompt were 2.6 times the corresponding odds when he gave a prompt to reflect on an ACTION.

Complex descriptions, too, were evidence of rich discourse; they provided layers of description or genuine insight about how students interpreted an experience or observation. ACTION prompts were strongly associated with this type of response, as nearly 49% of *Complex descriptions* were preceded by prompts to reflect on actions (and 27% of ACTION prompts were followed by *Complex descriptions*). According to the log-multiplicative model, the odds of eliciting a *Complex description* response versus a *Simple Description* when the tutor gave an ACTION prompt were 2.0 times the corresponding odds when he gave an APPLY prompt, and 1.6 times the corresponding odds when he gave a METACOGNITIVE prompt.

Revealing knowledge in transition

Before arriving at explicit statements of understanding, students can experience a phase of uncertainty during which their knowledge is in transition (Perry and Lewis 1999). If tutors can pay attention to indexes of uncertainty, they can opportunistically capitalize on them to promote development of new knowledge (Goldin-Meadow et al. 1993). With this in mind, we sought to determine which prompt types were most closely associated with students' expressions of indecision and uncertainty, as evidenced by *Impasses*, *Silence*, and asking *Questions*. For these odds ratios, it made sense to use *Simple Descriptions* as the comparison response; these responses were typically expressed with considerable confidence and typically described unambiguous phenomena. For similar reasons, it made sense to use ACTION prompts as the comparison prompt type. We conjectured that APPLY, EXPLANATION, and METACOGNITIVE prompts would be strongly associated with responses that revealed student uncertainty because these were likely to put the students in a position to reckon with or to admit what they did not know.

APPLY prompts were effective at eliciting responses that reflected student uncertainty. We found that the odds of yielding a *Question* response versus a *Simple Description* when the tutor gave an APPLY prompt were 14.3 times the corresponding odds when he gave a prompt to reflect on an ACTION. Substituting *Silence* responses for *Question* responses, the corresponding odds ratio is 16.4, and for *Impasses* the odds ratio becomes 4.0. EXPLAIN prompts, too, were strongly associated with *Questions*, *Silence*, and *Impasses*: the corresponding odds ratios were 4.7, 4.1, and 2.4, respectively. Finally, METACOGNITIVE prompts were effective at eliciting uncertainty as well. The odds of yielding a *Question* response versus a *Simple Description* when the tutor gave a METACOGNITIVE prompt were 8.2 times the corresponding odds when he gave a prompt to reflect on an ACTION. Substituting *Silence* responses for *Question* responses, the corresponding odds ratio was 7.3, and for *Impasses* the odds ratio was 3.3.

Discussion

Despite the reported successes of reflection in educational contexts (e.g., Cognition and Technology Group at Vanderbilt 1992), and long-standing recommendations to encourage reflection (e.g., Dewey 1916), research has shone scarce light on the dynamic process of reflection itself as a tool for learning (Davis 2003). Without a better understanding of these processes, it is difficult to give teachers guidance about exactly what they should be encouraging, or what results they will likely obtain if they actually encourage reflection. Thus, we set out to learn more about the process of reflection by focusing on turn-by-turn analyses of seven young students working with a tutor during sustained scientific inquiry into gear movement.

Before we provide our final words on how we understand the results and their implications, we offer a caveat: We began with a small number of participants so that we could closely analyze moment-by-moment interactions (Siegler and Jenkins 1989), but this small sample size prohibits us from generalizing far from the one tutor we observed. Likewise, some of our findings might be limited to the specific domain of gear mechanics. Even so, the fact that prompts to reflect could be so prevalent in a tutoring situation, with a diverse group of young students, and that the students were so responsive to the tutor's prompts, should inspire more research on promoting student reflection about scientific concepts.

In the remainder of this discussion, we hope to provide insight at two levels. First, we review what we attempted to learn by conducting this investigation and present what was generally obtained from our data and analyses. Second, we extend our conclusions by exploring our 2-dimensional model and, taking the model as a point of departure, by making conjectures about how students can develop scientific inquiry skills in the context of tutoring sessions.

What we wanted to learn and what we actually learned from this investigation

In this investigation, we aimed to explore five interrelated issues: (1) *How frequently* do prompts for reflection naturally occur in scientific, investigative tutoring situations? (2) What *types of prompts* does an expert tutor use to encourage student reflection and how do these prompt types differ substantively from each other? (3) How do *students respond* when provided with prompts for reflection? and (4) Is there a *relationship between the types of prompts received and the responses* students give?

Frequency of prompting students to reflect

Our results clearly point to how frequently and fluidly a tutor can elicit students to reflect on principles of gear movement. We found that the tutor in our investigation prompted for reflection, on average, once almost every 20s, which certainly had to impact student learning. Although in some contexts the high frequency of prompts might be a signal of tutor inexperience (Chi et al. 2008), given that our tutor had more than 10 years of experience in similar instructional settings, this is not necessarily a mark of inexperience. The tutor made frequent prompts to sustain students' attention to detail and to build a foundation for challenging reflective requests.

What types of prompts did the tutor use?

We were successful in reliably categorizing the tutor's prompts. The tutor provided a diversity of prompt types—seven types varying in substance and purpose—to each student in our investigation. This suggests that instructors have a range of possible strategies to evoke potential reflection from their students.

The tutor favored some prompts. Whereas low-level prompts privileged students' own observations, the higher level prompts often were an attempt to negotiate meaning around real world applications, explanations, and predictions. As Dillenbourg et al. (1996, p. 19) suggest:

> There is [a] type of negotiation that is common to any verbal interaction, and which takes place at the communicative, rather than the task, level: negotiation of meaning. The general idea is that the meaning of utterances in verbal interaction…is not something that is fixed by speakers and their utterances, but is rather something to be jointly constructed throughout the interaction by both speakers.

Through the variety of prompts used, the tutor in our study encouraged reflection as a means to negotiate meaning and to propel investigation. For example, although there are precise scientific terms for gear artifacts, gear motion, and the physical relationships between moving gears, the tutor worked through these ideas with the students by privileging their own language and schemas. In this way, the tutor joined the student as a partner in learning.

An additional factor contributed to the diversity of prompts: the influence of the students' contributions themselves. In their study on the role of observation of tutoring compared to actual participation, Chi et al. (2008) surmised that the tutor is not the determining factor in effective tutoring, but that it is the "contributions of the tutees themselves, and their interactions with the tutor, [that are] responsible for learning" (p. 337). Graesser et al. (1995) made a related point: that the effectiveness of a tutorial dialogue depends less on the tutor's "esoteric pedagogical strategies" (p. 496) than on sustaining interactive learning. In our study, the tutor was very responsive to students' contributions, thereby sustaining their engagement in a highly interactive dialogue.

How can we describe student responses?

We then turned to an examination of the effects of tutor prompts. We found that the students were sensitive to these prompts: They responded verbally to prompts 87% of the time. Clearly, prompting students to reflect will lead them to respond in discernable ways. The level of difficulty of the prompt mattered, as students were much more likely to provide high-level responses when given a high-level prompt than when given a low-level prompt. Thus, it appears that students are quite sensitive to the level of rigor in the way a tutor prompts them to reflect.

We identified 12 different categories of student responses; students demonstrated great variation in how they responded to prompts. This heterogeneity alone illustrates how prompts for reflection engaged students in complex discourse with the tutor. In turn, this complexity not only sustained students' inquiry over the course of a tutoring session, but also facilitated multidimensional reflection on scientific ideas.

To the extent that the tutor succeeded in joining the tutee as a partner in learning, the patterns of students' reflections paralleled the learning that occurs in classrooms that feature "true dialogue" (Lemke 1990; van Zee and Minstrell 1997). With low-level prompts, our tutor directed the student to observe critical details; he guided like a teacher. With thoughtful high-level prompts, the tutor requested deep explanations in the student's authentic language for mutual understanding; he assumed a collaborative learner role. During the session, the student became immersed in the role of a dynamic, respected partner, which, in turn, contributed to a flexible learning environment. Thus, students' high response rates and the wide variation of responses were manifested not only in the diversity and contextual appropriateness of tutor prompts, but also in a prevailing reflective atmosphere. In this sense, our findings are similar to those of Chi et al. (2001), who observed that tutoring is optimal when the student and tutor engage in a "collaborative, co-constructing way" (p. 41).

Relations between prompts and responses

Although we were struck by the sheer number of prompts by the tutor and by the responsiveness from the students, we were perhaps most impressed by the kinds of relationships we observed between prompt types and response types. The log-linear and log-multiplicative association modeling revealed a structure of associations that could be described along two dimensions: temporality and certainty. In the remainder of the discussion, we attend to the relations between prompts and responses that seemed robust from our modeling of the data and odds-ratio analyses. In particular, if the goal of a tutoring session is, at least in part, to reveal misconceptions and unstable knowledge and to help students revise and rebuild conceptually sound knowledge, we should pay attention to how the tutor reveals unstable knowledge and provokes the student to develop new knowledge.

Putting the results in context: two dimensions of reflective learning

The analyses and models that allowed us to visualize the relationship between prompts and responses yielded a 2-dimensional model. We inferred post-hoc the two dimensions to represent time and certainty, respectively. In considering time, we expect that prompting students to reflect on the past has different developmental outcomes than prompting reflection on the future, although both are valuable in constructing a sophisticated conceptual understanding of a physics, or other scientific, concept. Reflecting on the past has the potential to solidify for the student what he or she already knows (Bouillion 2007); reflecting on the future has the potential to engage the student in processes of hypothesis formation, prediction, and scientific inquiry, more generally. Thus, prompts throughout the time continuum positively affect learning, albeit in different ways.

The second dimension, certainty, reveals to the tutor what conceptual material remains elusive to the student (Graham and Perry 1993; Perry and Lewis 1999) or what the student has already mastered (e.g., responds with a *Complex Description*). Thus, the student's displayed uncertainty works to make obvious to both the student and tutor that the student may be facing a deficit or uncertainty in understanding. By prompting the student to display uncertainty, the tutor can hone in on those aspects of the concept with which the student is struggling (Alibali et al. 1997)—but with discretion. Impasses, under the right conditions, can motivate students to take a more active role in forming better understandings (Sánchez et al. 2009; VanLehn et al. 2003).

Not only do the tutor's prompts provide opportunities to reveal students' misunderstandings and developing knowledge, but also to provide a bridge to new knowledge (see, e.g., Bransford et al. 2000). If the student is put in a position to ask a question, the tutor can respond with information to satisfy the student's intellectual concern. If the student has an impasse or silence, the tutor has the cue to step in and provide guidance to help the student out of the confused state. By giving APPLY, EXPLAIN, or METACOGNITIVE prompts, the tutor can help to move along a student's understanding of gear movement.

Before students are even able to articulate new knowledge, sometimes they first need to make observations and describe what they observe, using these components to build their new knowledge. By engaging students in this practice, tutors force students to reckon with aspects of the problem that heretofore may have gone unnoticed and, even if noticed, had not been integrated into their conceptual knowledge (Chi et al. 1994; Hatano and Inagaki 1991). For example, merely acknowledging that adjacent gears turn in opposite directions enables students to come to terms with basic and then more sophisticated understandings of gear movement (Perry and Elder 1997). These cognitive processes are not necessarily explicit. As the literature on scaffolding has demonstrated, children's problem solving often has a "disordered and deep structure" (Wood et al. 1976, p. 97) that is not readily apparent, yet is responsive to moment-to-moment engagement with an expert.

By emphasizing the positive aspects of uncovering uncertainty, we do not mean to undermine the importance of teachers or tutors provoking students to explain what they know. Requesting students to explain, whether they actually can explain or not, enables the tutor to both expose what is not yet understood and also to allow students to articulate their new knowledge. When students produce an explanation (e.g., *Complex Description*), and it is shared with an expert, the student builds an ownership of that new knowledge (Pressley et al. 1992). Thus, the act of requesting conceptual information from students puts both tutor and student in the position to expand and solidify developing conceptual knowledge.

The two dimensions uncovered by our analyses work not in isolation but synergistically. As an example, we take a brief look at how REAL WORLD prompts functioned in the tutoring episodes we observed. REAL WORLD prompts asked students to reflect on past events, but also yielded relatively unstable knowledge. Although this may seem surprising at first, on further consideration, we realized that when students were asked to tie the scientific problem to what they already knew, they had a hard time making clear connections to their everyday experiences. As Bouillion (2007) has noted, having students make these types of connections is often critical for students to build meaningful scientific concepts. This provides grounds for claiming that having instructors prompt students to make connections first disrupts their naive conceptions (Reiner et al. 2000) and then opens the way for students to make better sense of their everyday experiences in terms of scientific principles. Prompting students to reflect on real world experiences may have motivational benefits as well; as recent experimental evidence suggests, prompting students to make connections between science learning and their own lives boosts interest in science and course grades—particularly among students with low expectations for success (Hulleman and Harackiewicz 2009).

Instructors have much power at their hands when they take charge of children's learning. Our investigation suggests that it is possible, and conceivably quite natural, for instructors to entice young students to reflect on their scientific problem solving. We have provided some evidence for how this process of reflection operates in a tutoring situation. Our desire is that researchers will be inspired to understand this phenomenon better and that instructors will be inspired to learn more about how they can use prompts to engage their students in productive reflection. Ultimately, we hope that by shedding light on this process, students will be the beneficiaries, with improved understanding of scientific concepts and enhanced learning outcomes.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley.

Alibali, M. W. (2005). Mechanisms of change in the development of mathematical reasoning. In R. Kail (Ed.), *Advances in child development and behavior* (Vol. 33, pp. 79–123). New York: Academic Press.

Alibali, M. W., Flevares, L. M., & Goldin-Meadow, S. (1997). Assessing knowledge conveyed in gesture: Do teachers have the upper hand? *Journal of Educational Psychology, 89*, 183–193.

Alibali, M. W., & Goldin-Meadow, S. (1993). Gesture-speech mismatch and mechanisms of learning: What the hands reveal about a child's state of mind. *Cognitive Psychology, 25*, 468–523.

Becker, M. P. (1989). Models for the analysis of association in multivariate contingency tables. *Journal of the American Statistical Association, 84*, 1014–1019.

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. New York: McKay.

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.

Bouillion, L. (2007). Students' conceptual practices in science education. *Cultural Studies of Science Education, 2*, 205–223.

Bransford, J., Brown, A. L., & Cocking, R. R. (2000). *How people learn*. Washington, DC: National Academy Press.

Broaders, S. C., Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2007). Making children gesture brings out implicit knowledge and leads to learning. *Journal of Experimental Psychology: General, 136*, 539–550.

Brown, A. L., & Campione, J. C. (1990). Interactive learning environments and the teaching of science and mathematics. In M. Gardner, et al. (Eds.), *Toward a scientific practice of science education*. Mahwah, NJ: Lawrence Erlbaum Associates.

Bullock, M., Gelman, R., & Baillargeon, R. (1982). The development of causal reasoning. In W. J. Friedman (Ed.), *The developmental psychology of time* (pp. 209–254). New York: Academic Press.

Chambers, J. M., Carbonaro, M., & Murray, H. (2008). Developing conceptual understanding of mechanical advantage through the use of Lego robotic technology. *Australasian Journal of Educational Technology, 24*, 387–401.

Chi, M., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*, 145–182.

Chi, M. T. H., de Leeuw, N., Chiu, M.-H., & Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*, 439–477.

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

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

Chinn, C. A., & Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools: A theoretical framework for evaluating inquiry tasks. *Science Education, 86*, 175–218.

Cognition and Technology Group at Vanderbilt. (1992). The Jasper Series as an example of anchored instruction: Theory, program description, and assessment data. *Educational Psychologist, 27*, 291–315.

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

Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2008). Gesturing makes learning last. *Cognition, 106*, 1047–1058.

Davis, E. A. (2003). Prompting middle school science students for productive reflection: Generic and directed prompts. *The Journal of the Learning Sciences, 12*, 91–142.

Dewey, J. (1916). *Democracy and education*. New York: MacMillan.

Dillenbourg, P., Baker, M., Blaye, A., & O'Malley, C. (1996). The evolution of research on collaborative learning. In E. Spada & P. Reiman (Eds.), *Learning in humans and machine: Towards an interdisciplinary learning science* (pp. 189–211). Oxford: Elsevier.

diSessa, A. (1993). Toward an epistemology of physics. *Cognition and Instruction, 10*, 105–225.

Goldin-Meadow, S., Alibali, M. W., & Church, R. B. (1993). Transitions in concept acquisition: Using the hand to read the mind. *Psychological Review, 100*, 279–297.

Goodman, L. A. (1981). Association models and bivariate normal for contingency tables with ordered categories. *Biometrika, 68*, 347–355.

Graesser, A. C. (1992). *Questioning mechanisms during complex learning*. Memphis, TN: Memphis State University (ERIC Document Reproduction Service No. ED 350 306).

Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal, 31*, 104–137.

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.

Graham, T., & Perry, M. (1993). Indexing transitional knowledge. *Developmental Psychology, 29*, 779–788.

Hatano, G., & Inagaki, K. (1991). Sharing cognition through collective comprehension activity. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 331–348). Washington, DC: American Psychological Association.

Hosenfeld, B., van der Maas, H. L. J., & van den Boom, D. C. (1997). Indicators of discontinuous change in the development of analogical reasoning. *Journal of Experimental Child Psychology, 64*, 367–395.

Hulleman, C. S., & Harackiewicz, J. M. (2009). Promoting interest and performance in high school science classes. *Science, 326*, 1410–1412.

Krajcik, J., Blumenfeld, P., Marx, R., Bass, K. M., Fredericks, J., & Soloway, E. (1998). Middle school students' initial attempts at inquiry in project-based science classrooms. *The Journal of the Learning Sciences, 7*, 313–350.

Lehrer, R., & Schauble, L. (1998). Reasoning about structure and function: Children's conceptions of gears. *Journal of Research in Science Teaching, 35*, 3–25.

Lemke, J. L. (1990). *Talking science: Language learning, and values*. Norwood, NJ: Ablex.

Linn, M. C. (1995). Designing computer learning environments for engineering and computer science: The scaffolded knowledge integration framework. *Journal of Science Education and Technology, 4*, 103–126.

Loh, B., Reiser, B. J., Radinsky, J., Edelson, D. C., Gomez, L. M., & Marshall, S. (2001). Developing reflective inquiry practices: A case study of software, the teacher, and students. In K. Crowley, C. D. Schunn, & T. Okada (Eds.), *Designing for science: Implications from everyday, classroom, and professional settings* (pp. 279–323). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

May, D. B., & Etkina, E. (2002). College physics students' epistemological self-reflection and its relationship to conceptual learning. *American Journal of Physics, 70*, 1249–1258.

McNeill, K. L., Lizotte, D. J., Krajcik, J., & Marx, R. W. (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. *The Journal of the Learning Sciences, 15*(2), 153–191.

Metz, K. E. (1991). Development of explanation: Incremental and fundamental change in children's physics knowledge. *Journal of Research in Science Teaching, 28*, 785–797.

Mohan, M. (1972). *Peer tutoring as a technique for teaching the unmotivated*. Fredonia, NY: State University of New York, Teacher Education Research Center (ERIC Document Reproduction Service No. ED 061 154).

Ozgungor, S., & Guthrie, J. T. (2004). Interactions among elaborative interrogation, knowledge, and interest in the process of constructing knowledge from text. *Journal of Educational Psychology, 96*, 437–443.

Perry, M., Church, R. B., & Goldin-Meadow, S. (1988). Transitional knowledge in the acquisition of concepts. *Cognitive Development, 3*, 359–400.

Perry, M., & Elder, A. D. (1997). Knowledge in transition: Adults' developing understanding of a principle of physical causality. *Cognitive Development, 12*, 131–157.

Perry, M., & Lewis, J. (1999). Verbal imprecision as a tool for understanding knowledge in transition. *Developmental Psychology, 35*, 749–759.

Person, N. K., Graesser, A. C., Magliano, J. P., & Kreuz, R. J. (1994). Inferring what the student knows in one-to-one tutoring: The role of student questions and answers. *Learning and Individual Differences, 6*, 205–229.

Piaget, J. (1964/1967). *Six psychological studies*. New York: Vintage Books.

Pressley, M., Wood, E., Woloshyn, V. E., Martin, V., King, A., & Menke, D. (1992). Encouraging mindful use of prior knowledge: Attempting to construct explanatory answers facilitates learning. *Educational Psychologist, 27*, 91–109.

Putnam, R. T. (1987). Structuring and adjusting content for students: A study of live and simulated tutoring of addition. *American Educational Research Journal, 24*, 13–48.

Recker, M., & Pirolli, P. (1995). Modeling individual differences in students' learning strategies. *The Journal of the Learning Sciences, 4*, 1–38.

Reiner, M., Slotta, J. D., Chi, M. T. H., & Resnick, L. B. (2000). New physics reasoning: A commitment to substance-based conceptions. *Cognition and Instruction, 18*, 1–34.

Reiser, B. J. (2004). Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *The Journal of the Learning Sciences, 13*, 273–304.

Rubie-Davies, C. M. (2007). Classroom interactions: Exploring the practices of high- and low-expectation teachers. *British Journal of Educational Psychology, 77*, 289–306.

Sánchez, E., García-Rodicio, H., & Acuña, S. R. (2009). Are instructional explanations more effective in the context of an impasse? *Instructional Science, 37*, 537–563.

Sandoval, W. A., & Reiser, B. J. (2004). Explanation-driven inquiry: Integrating conceptual and epistemic scaffolds for scientific inquiry. *Science Education, 88*, 345–372.

Schauble, L., Glaser, R., Duschl, R. A., Schulze, S., & John, J. (1995). Students' understanding of the objectives and procedures of experimentation in the science classroom. *The Journal of the Learning Sciences, 4*, 131–166.

Schon, D. A. (1992). The theory of inquiry: Dewey's legacy to education. *Curriculum Inquiry, 22*, 119–139.

Schraw, G., Crippen, K. J., & Hartley, K. (2006). Promoting self-regulation in science education: Metacognition as part of a broader perspective on learning. *Research in Science Education, 36*, 111–139.

Siegler, R. S., & Jenkins, E. A. (1989). *How children discover new strategies*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Simpson, E. H. (1949). Measurement of diversity. *Nature, 163*, 688.

Songer, N. B., & Linn, M. C. (1991). How do students' views of science influence knowledge integration? *Journal of Research in Science Teaching, 28*, 761–784.

van Zee, E., & Minstrell, J. (1997). Using questioning to guide student thinking. *The Journal of the Learning Sciences, 6*, 227–269.

VanLehn, K., Siler, S., Murray, R. C., Yamauchi, T., & Baggett, W. B. (2003). Why do only some events cause learning during human tutoring? *Cognition and Instruction, 21*(3), 209–249.

Vermunt, J. K. (1997). LEM: *A general program for the analysis of categorical data*. [Computer software manual]. Tilburg: Tilburg University. Retrieved from http://www.uvt.nl/faculteiten/fsw/organisatie/departementen/mto/software2.html.

Vygotsky, L. S. (1962). *Thought and language*. Cambridge, MA: The MIT Press.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

White, B. Y., & Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction, 16*, 3–118.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17*(2), 89–100.