(In Press) Cognition and Instruction

CAN TUTORS MONITOR STUDENTS' UNDERSTANDING ACCURATELY?

Michelene T. H. Chi, Stephanie A. Siler, Heisawn Jeong*

Learning Research and Development Center

University of Pittsburgh

Chi@pitt.edu

* Now at Hallym University, Chun Chon, Kangwon-do, Korea. Funding for this research is provided in part by the Spencer Foundation to the first author, and in part by the National Science Foundation, Grant Number NSF (LIS): 9720359, to the Center for Interdisciplinary Research on Constructive Learning Environment (CIRCLE). We are grateful for comments provided by Robert Hausmann and Kurt VanLehn. Copies for this paper may be requested from Micki Chi, or downloaded from the WEB: www.pitt.edu/~Chi

Abstract

Students learn more and gain greater understanding from one-to-one tutoring. The preferred explanation has been that the tutors' pedagogical skills are responsible for the learning gains. Pedagogical skills involve skillful execution of tactics, such as giving explanations and feedback, or selecting the appropriate problems or questions to ask the students. Skillful execution of these pedagogical skills requires that they are adaptive and tailored to the individual students' understanding. In order to be adaptive, the tutors must be able to monitor students' understanding accurately, so that they know how and when to deliver the explanations, feedback, and questions. Before exploring whether in fact tutoring effectiveness can be attributed to tutors' pedagogical skills, we must first ascertain the accuracy with which tutors monitor their students' understanding. This paper thus investigated monitoring accuracy from both the tutors' and the students' perspectives. By coding and re-coding some data collected in a previous study, the paper showed that tutors could only *assess* students' *normative* understanding from the perspective of the tutors' knowledge, but tutors were dismal at *diagnosing* the students' *alternative* understanding from the perspective of the students' knowledge.

Introduction

Human one-to-one tutoring has been shown to be a very effective form of instruction. The average student in a tutoring situation achieved a performance gain ranging from 0.4 to 2.3 standard deviations above the average student in a traditional transmission-based classroom (Bloom, 1984; Cohen, Kulik, & Kulik, 1982). A meta-analysis of around 100 studies showed that in educational settings, everyday tutors are often peers and other paraprofessionals with little experience in tutoring (Cohen, et al., 1982), although they are almost always knowledgeable about the content domain, such as algebra. Nevertheless, students benefit and learn from being tutored even when the tutors are inexperienced in the skills of tutoring.

Despite the lack of tutoring expertise, there's a tendency in the literature to attribute the overall effectiveness of tutoring to the tutors' pedagogical skills (Chi, Siler, Jeong, Yamauchi & Hausmann, 2001; Collins & Stevens, 1982; Merrill, Reiser, Ranney & Trafton, 1992; McArthur, Stasz & Zmuidzinas, 1990). Pedagogical skills refer to a variety of tutoring tactics, such as selecting the next problem for the students to work on, giving explanations and feedback, timing the feedback, eliciting and prompting for explanations, scaffolding, asking comprehension gauging questions, and so forth. The majority of studies on tutoring do show that tutors undertake these kinds of pedagogical tactics; moreover, tutors tend to dominate the tutoring sessions, such as by initiating and terminating the conversations (Chi, et al. 2001; Graesser, Person, & Magliano, 1995). However, the mere fact that tutors tend to dominate the tutoring sessions and undertake numerous tutoring moves does not necessarily lead to the conclusion that their tutoring tactics are skillfully executed. For example, do they select the *appropriate* next problems for the students to work on or do they simply pick the next "logical" problem, from the standpoint of a curriculum script; do they give explanations *tailored* to the students' needs and

misunderstanding, or do they simply give the standard explanations for a specific misunderstanding; do they give feedback at the *right moment* or do they give it whenever overt misunderstanding occurs (Cho, Michael, Rovick & Evens, 2000)? In short, are these tutoring tactics executed in a skillful way that maximizes students' learning? In other words, are tutors in fact adaptive?

The implicit prevailing view, fostered by the influential work of Collins and Stevens (1982; Stevens & Collins, 1977) and consistent with the general behavioral evidence (in which tutors dominate the dialogues), takes a tutor-centered perspective, in that it assumes that the tutors (whether they are experienced or not) are responsible for tutoring effectiveness, in the way they craft and customize their explanations to the needs of the students. Thus, a tutor-centered view assumes that the tutors' tactics are undertaken in an adaptive way, based on the students' ongoing understanding. (See Chi et al., 2001, for a more comprehensive review of the different hypotheses for tutoring effectiveness.)

However, in order for tutors to be adaptive in the skillful execution of any of the tutoring tactics, the tutors must monitor the students' mis- or incorrect understanding accurately. (The terms incorrect and misunderstanding will be used interchangeably.) Thus, in some sense, a critical prerequisite skill is a monitoring one. In order to test the assumption that it is the adaptive nature of the tutors' pedagogical skill that is responsible for tutoring effectiveness, this paper addresses the prerequisite question of how accurately can tutors monitor the status of their students' misunderstanding. Thus, the goal of this paper is to see the accuracy with which novice tutors succeeded in monitoring their students' misunderstanding.

In order to address the issue of whether or not tutors customize their tutoring to the students' misunderstanding, we need to first define and differentiate two forms of incorrect

understanding: "normative" versus "alternative." These differences are important and will be elaborated below (see Table 1, left column).

Normative versus Alternative Understanding

The study described in this paper examines learning of conceptual knowledge about the human circulatory system. We define *correct* conceptual knowledge to be the normative or scientific knowledge (these terms will be used interchangeably) that is addressed in the text of our study. However, in a tutoring context, correct knowledge can also be the tutors' expectations, which may deviate somewhat from the normative knowledge.

Conceptual knowledge can be analyzed at two different grain sizes: beliefs (at the local proposition level) or mental models (at the global structural level). Beliefs can be analyzed in an isolated piecemeal way, whereas mental models can be analyzed in an integrated coherent way. We assume that as students learn, they are building mental models to represent the information they are learning. Presumably, a mental model represents some sort of integrated structure in that the various beliefs embodied in a mental model are organized in some meaningful way. Thus, a correct mental model (such as of the human circulatory system) would be one that is analogous to the actual physical (and scientifically correct) model of the human circulatory system, in terms of its components, their relationship, and how blood flows through the components (Gentner & Stevens, 1983). Students' knowledge can be captured at the mental model level by analyzing their explanation protocols for critical features (see Chi, de Leeuw, Chiu & LaVancher, 1994, for a procedure of capturing students' mental models of the circulatory system).

Analyzing knowledge at two different grain sizes serves a useful distinction because each grain size captures a different aspect of understanding. Because beliefs are often assessed in a

piecemeal isolated way, knowing them generally correlates with shallow understanding (Chi et al., 2001). On the other hand, because mental models are often analyzed in an integrated coherent way (Gentner & Stevens, 1983), the status of students' mental models often reflects their deep understanding (Chi, et al., 1994; Chi, 2000). The importance of this distinction will be further shown later. The next sections differentiate incorrect "normative" from "alternative" understanding at both the belief and the mental model levels.

Table 1.
Two forms of incorrect understanding: Normative versus alternative

	BELIEFS	MENTAL MODELS
NORMATIVE UNDERSTANDING	Contradictory	Fragmented
ALTERNATIVE UNDERSTANDING	False (or True)	Flawed
ROBUST ALTERNATIVE UNDERSTANDING	Misconceptions	Incommensurate Theories

Contradictory Versus False Beliefs

We define an incorrect belief to be *contradictory* if a correct proposition that is either explicitly or implicitly stated in a text contradicts it. For example, a contradictory belief about the human circulatory system would be that *blood goes to the various body parts after it leaves the right ventricle.* Such a belief is *contradictory* in the sense that a text passage on the human circulatory system might contain sentences that directly refute it, such as stating that *Blood goes to the lungs after it leaves the right ventricle.* A *contradictory* belief can also be refuted in a text indirectly. For example, 8th graders often believe that *all blood vessels have valves.* Although a

text may never explicitly state that *Only some vessels have valves,* the fact that a text describes valves only in the context of veins and never in the context of arteries, would constitute indirect refutation of a student's contradictory belief.

Contradictory beliefs that existed prior to instruction can be readily corrected once a student has read a relevant text passage. In a previous study in which we identified several prior contradictory beliefs among 8th graders, 92% of these contradictory beliefs were revised or removed after the students read the text we used in our studies, which never mentions valves in the context of the arteries but only in the context of veins. Thus, such indirect refutation succeeded in revising the majority of students' initial contradictory beliefs (de Leeuw, 1993; also reported in Chi & Roscoe, 2002). In short, a *contradictory* belief is an incorrect belief that a scientific proposition contradicts, and it can be readily removed after reading a text that refutes it, either directly or indirectly. (Of course, contradictory beliefs can also be refuted by other sources, such as the media and other personal experiences. But our data pertained only to text refutations.) By our definition then, having *contradictory* beliefs is a form of incorrect "normative" understanding, at the proposition level.

On the other hand, *false* beliefs are incorrect beliefs that are presumably not addressed by any texts that the students have encountered. Because they were never addressed by the text or by the tutors in a specific instructional context (see row 2, Table 1), they may not be removed even after instruction. That is, students with a *false* belief will often not be exposed to its corresponding correct fact. For example, students sometimes think that *veins are like nerves, they transmit signals from the brain,* or that *blood is heavy* when neither transmitting nor heaviness are properties of veins and blood that are likely to be addressed in textbooks. In our prior study, none of the students' initial false beliefs were removed after instruction (i.e., reading the text),

since they were not explicitly refuted by the text, whereas, as mentioned above, 92% of the contradictory beliefs were correctly revised (Chi & Roscoe, 2002; de Leeuw, 1993). Because false beliefs are constructed by the students and not typically addressed by any texts, false beliefs often persist and represent a form of "alternative" misunderstanding at the proposition level. Thus, the differentiation between *contradictory* and *false* beliefs may be an important distinction. (Note that since alternative understanding is constructed by the students, occasionally students can construct beliefs that are correct; these will be referred to later as *true* alternative beliefs.)

By our preceding definitions and the results of de Leeuw (1993), false beliefs are not in principle difficult to remove; they linger because they are not addressed by the text. However, there are alternative beliefs that persist even when they are both directly and indirectly confronted by instruction and refutation, to be referred to here as *misconceptions*. An example would be students' naive belief that *heat is made of hot molecules.* These kinds of *misconceptions* resist correction for a different set of reasons that cannot be addressed here (instead, see Chi, in press; and Chi & Roscoe, 2002). They are included in the bottom row of Table 1 for completeness, under the category "robust alternative understanding."

Fragmented Versus Flawed Mental Models

There can also be two ways to conceive of students' incorrect understanding in the context of mental models: *fragmented* versus *flawed* (see Table 1 again, right column). A *fragmented* mental model is one that is normatively incorrect (in the sense of either missing some components of the correct models, missing some details of the components, missing or weakly linked relations among the components, etc.), as compared to the correct scientific model. Being *fragmented* also means that the model cannot be used to make systematic

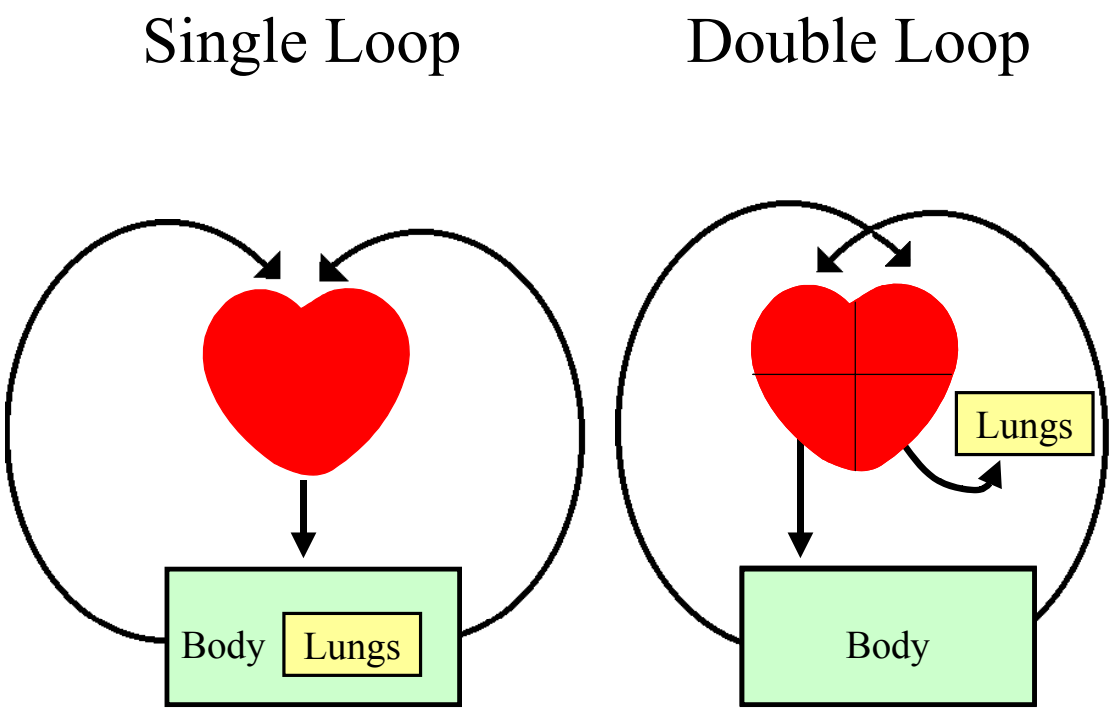
predictions or generate sensible and consistent explanations. On the other hand, a *flawed* mental model would be one in which the relationships among the components are incorrect, but they are nevertheless coherently organized. That is, students can use their flawed mental models to generate systematic predictions and explanations, and do so consistently. A typical flawed mental model of the circulatory system is a "single loop with no lungs" model, in which the students think blood simply goes from the heart to different parts of the body (including the lungs as a part of the body). The left side of Figure 1 depicts an idealized "single loop with no lungs" model. Such a model does not recognize the important role lungs play in oxygenation of the blood; it assumes that blood is produced in the heart and it is in the heart that blood gets its oxygen. Below is a sample explanation given by a student (J) when asked to define the circulatory system:

> *"The process of the way the blood goes in your body, you know, like goes through your body and then to your heart and then starts the process over from where it starts… It starts on one part of the heart and goes through the body and then comes back to the heart again and the process starts over again."*

A "single loop" model, thus, is an "alternative" model in that it makes three assumptions that are fundamentally different from the scientifically correct "double loop" model. That is, students with this model assume that (1) the heart (rather than the lungs) is the source of oxygen, that (2) blood goes to the lungs to deliver oxygen (rather than to exchange oxygen and carbon dioxide), and that (3) there is a single (rather than a double) loop. All "single loop" models share these three fundamental assumptions.

If the same student (J) is further pressed to include lungs, hands and feet in his explanation, then he would modify his explanation, but it is still compatible with his basic single loop model, in that lungs are just another body part to which blood has to travel, as:

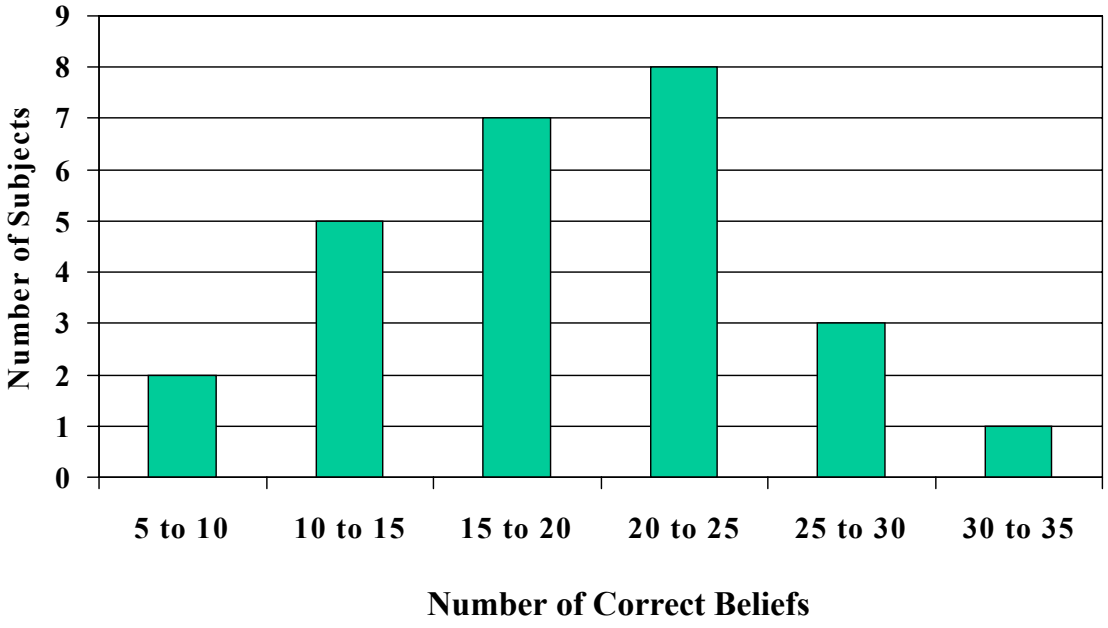
"Okay. It goes in the heart, to the hands…to the hands…up into the brain… and from the brain back to the heart. Um…from the heart to the lungs… from the lungs to the brain and then from the brain back to the heart. The heart… to the feet and then from the feet back, I mean up to the brain." Thus, students' flawed mental models are *coherent* in the sense that they can give consistent and systematic explanations based on their flawed models (Chi., 2000; Chi, et al., 1994; and Vosniadou & Brewer, 1992).



Furthermore, students with a single loop model appear not to be aware that they have a flawed model (Chi, 2000; Chi & Roscoe, 2002), because our subjective impression is that they generate explanations to questions confidently without any hesitations or confusion, and thus, their explanations based on the single loop model seem right to them. Thus, we define a mental model to be *flawed* in the following senses: (a) that it entails a distinct set of assumptions (as compared to a correct mental model), (b) it is coherent (as defined above), and (c) students appear to be unaware that their flawed mental models are incorrect. Having a flawed mental model represents a form of "alternative" incorrect understanding.

In Chi, et al. (1994), using protocol analyses, we were able to identify five types of flawed mental models. Over half of the students' initial pre-instruction mental models were of the "single loop" type. Even though all flawed "single loop" mental models share the same three basic assumptions, there can nevertheless be several variations in the details as well as in the number of correct beliefs of which each model is composed. That is, a flawed mental model is not necessarily one that is made up of all incorrect beliefs, whether contradictory or false ones. The critical factor of a flawed mental model is how the beliefs are organized, not how many of them are incorrect. In data collected by Jeong (1998) and reported in Chi and Roscoe (2002), 40 college students' pre- and post-instruction conceptions of the circulatory system were coded at both the belief and the mental model levels. Twenty-six of these 40 students held a single loop model, and yet the number of correct beliefs embedded in their single loop models ranged from 5 to 35 (see Figure 2). For example, although 14 students had between 5-20 correct beliefs, 12 students had between 20-35 correct beliefs. Thus, flawed mental models (such as a single loop model) can often be composed of many correct beliefs (as many as 30-35). Therefore, the number of correct beliefs does not determine the correctness of a mental model. This variability

in the number of correct beliefs per model validates our assumption of the need to differentiate knowing at the belief level versus the mental model level. It suggests that isolated assessment of the number of correct beliefs at the local proposition level is shallow and misleading, because it does not reflect deep understanding assessed at the integrated mental model level[1].



To summarize, *contradictory* beliefs and *fragmented* mental models both represent incorrect "normative" understanding as determined by comparing them with the scientifically correct beliefs and the scientifically correct mental models. *False* beliefs and *flawed* mental models, on the other hand, represent students' incorrect "alternative" understanding that may not correspond to the scientifically correct normative knowledge. The question pertinent to this paper is the extent to which tutors are aware of students' alternative understanding (to be assessed in the context of their false beliefs and flawed mental models). Again, not addressed by

[1] This result also has serious implication regarding the utility of corrective feedback that is given on the basis of piecemeal assessment at the proposition level. Such local feedback may or may not guarantee that students' global mental models will necessarily be correctly revised.

this paper are the *misconceptions* and *incommensurate theories* in which these alternative flawed mental models are embedded. These issues of "robust alternative" understanding are addressed in Chi (in press) and Chi and Roscoe (2002).

Two Forms of Monitoring

We propose that monitoring can be differentiated into two forms, which we will call *assessment* and *diagnosis*. We define *assessment* to be a form of monitoring that is evaluated from the perspective of the tutors' scientific (or normative) knowledge, whereas *diagnosis* is a form of monitoring that forces the tutor to take the perspective of the students' "alternative" knowledge. More specifically, assessing means that the tutors judge students' "normative" understanding on the basis of the correctness or incorrectness of the students' knowledge as compared to the scientific knowledge. One way to think about this is to conceptualize normative knowledge as represented in a correct template of nodes and relations. Assessing from a normative reference frame can be metaphorically conceptualized as overlaying the normative template over a student's utterances to see the degree of match in the concepts and relations. Thus, assessment evaluates the degree of incorrectness in a student's normative understanding, whereas diagnosis evaluates a student's alternative understanding.

The evidence in the literature so far suggests that tutors seldom give customized feedback that is based on diagnoses of students' alternative understanding. The strongest evidence comes from Putnam's (1987) study of 6 experienced classroom teachers tutoring first graders in multicolumn addition. His research goal was to see the extent to which tutors formulated an accurate model of a student's misconceptions and then proceed to remediate the misconceptions. Putnam (1987) determined whether or not tutors were diagnosing by analyzing which problems

the tutors chose to have the students work on next, in order to find out whether the tutors selected the problems to adapt to the students' understanding. On the basis of the sequence of problems selected, there was little evidence to support a diagnose-and-remediate pattern (Putnam, 1987). Instead of tailoring the selection of the to-be-solved problems to the student's understanding, the tutors tended to move through a linear and uniform sequence of problem types similar to the sequence presented in a standard curriculum script. Moreover, only 7% of students' errors were followed by diagnostic questions, such as "What were you thinking of when you added that?".

Likewise, in a microgenetic analysis of one professional tutor tutoring a student in solving three physics problems, the same pattern of tutoring was found in Chi (1996). Out of 41 tutorial actions coded for one of the problems, there were six occasions when the tutee clearly manifested confusion. In five of these six cases, this experienced tutor simply ignored the student's confusion, and instead, focused on following his own plan of teaching the student some specific procedure, such as how to draw all the forces on a diagram. Although getting such a procedure right is essential for learning how to solve this kind of problems, it was not the difficulty that the student was facing at the time. Thus, the results of this microgenetic analysis, showing that the tutor not only failed to address the student's difficulty on five out of the six occasions, but moreover, the tutor focused instead on what he assumed the student needed to learn, also suggest that even experienced professional tutors may not be effective at diagnosing students' alternative understandings.

Consistent with Putnam's finding, McArthur, Stasz and Zmuidzinas (1990) also suggested that there was no evidence that their tutors (who were teachers that had won awards for their excellence in teaching) made inferences about students' alternative knowledge. Instead, their tutors appeared to be interested only in students' self-monitoring of comprehension, by

asking questions such as "Do you understand?", "Have you heard of something called the additive inverse?", and "Have you done these before ever?" (p. 211). These kinds of questions are not diagnostic of students' alternative understandings; rather, they are comprehension-gauging questions (Chi et al., 2001) that ask the students *to determine for themselves* whether or not they understand. Finally, Graesser et al. (1995) also found that only 8% of the tutorial interactions were devoted to either attempts at or correction of students' misconceptions (p. 514**),** essentially the same as the 7% diagnostic questions reported by Putnam (1987). Graesser et al.'s tutors were either graduate students knowledgeable in the content domain, or experienced tutors with 9 hours of prior tutoring experiences.

Taken together, the evidence provided by Putnam (1987), Chi (1996), McArthur et al. (1990), and Graesser et al. (1995), all suggests that both experienced tutors and novice tutors (who were experienced teachers), do try to find out what information about the domain their students know from the normative (but not the alternative) perspective. However, such evidence is only suggestive either because it is based on a coarse-grained analysis of the next problems tutors chose, or because it is inferred from the moves that the tutors did not express (i.e., the lack of many specific diagnostic moves), and not based on any direct measures of what tutors in fact know about students' alternative knowledge. Generally it is very difficult to determine whether and how successfully tutors are monitoring students' understanding by examining the tutoring protocols alone. It is difficult because there are very few explicit monitoring comments in tutoring protocols, other than comprehension gauging questions that ask the *students themselves* to reflect whether or not they have understood the material. This study attempts to circumvent this problem by a methodology that was designed explicitly to ascertain the tutors' monitoring skills directly, by asking tutors in the midst of tutoring what their conceptions of their students'

knowledge are, as well as a fine-grained reanalysis that focused on tutors' ability to detect students' incorrect knowledge while tutoring.

A Naturalistic Tutoring Study

The goal of this paper is to provide more fine-grained and direct evidence, using a larger sample of novice (but domain-knowledgeable) tutors, to examine the extent to which they can accurately monitor students' understanding. There are four reasons for focusing on novice tutors. The most important one is that the majority of everyday tutors are novices (Fitz-Gibbon, 1977); even so, the tutees gained in learning from being tutored by them (Cohen et al., 1982). Second, the four previous studies cited above have already examined expert tutors (they were either experienced teachers or experienced tutors), so that it is useful to see how the results of inexperienced tutors compare with expert tutors. Third, because we still have little understanding of what tutoring skills are, it is difficult to know how to select a uniform group of expert tutors, other than from the amount of experience they have had tutoring. Finally, in order to gather a sample of a dozen or so tutors on the same topic, we had to resort to finding novice (but domain-knowledgeable) tutors.

The protocol data for this paper were collected in a previously published study on naturalistic tutoring (Study 1 reported in Chi et al., 2001). The analyses reported here consist of coding a portion of the protocol data that was not analyzed previously in the Chi et al. (2001) study, as well as a new recoding of the entire protocols for the purpose of this paper.

Method

Participants

Tutors were 11 college students (6 female nursing students, 1 male nursing student, and 4 male biology majors at the University of Pittsburgh) who were knowledgeable in the domain of the human circulatory system, but they were unskilled tutors. That is, they had neither tutoring experiences nor tutoring training. The students were 22 eighth-graders, who were recruited from various schools in the Pittsburgh area (both private and public schools). They were recruited either by having the teachers distribute letters for parents to contact us and give consent, or through a community-tutoring center. The students were compensated for their time at a rate of $6.00 per hour.

Materials

A passage about the human circulatory system, consisting of 86 sentences (see Appendix A in Chi et al., 2001), was taken from a well-written and popular biology text (Towle, 1989). Each sentence of the passage was printed on a separate sheet of paper (contained in a three-ring binder) so that both the tutors and students could know which sentence they were discussing. Pre-tests and post-tests consisted of terms and questions, and are presented in Appendix B, in Chi et al. (2001).

Design and Procedure

All the tutors were told to tutor the 11 students "naturally", in whatever manner they liked, except that they were asked to include motivating comments and avoid lecturing at the

students. The tutors were encouraged to limit the tutoring to no more than two hours and no less than one-and-a-half hours.

The study consisted of three sessions given in sequence - the pre-test session, the tutoring session, and the post-test session. In the pre-test and post-test sessions, students defined terms, answered questions, and drew the path of blood in circulation on an outline of the human body. The amount of learning, as assessed by the pre- and post-test results, is reported in Chi et al. (2001).

Before the tutoring session began, the tutors and the tutees were both required to read the entire passage once. During the tutoring session, the tutors and the tutees generally read the sentence on each page and engaged in a tutorial dialogue about each sentence. The tutoring dialogues were interrupted after the 24th and the 76th sentence of the text, when both the students and tutors were separately asked to draw the blood path of the circulatory system on a piece of paper which had an outline of the human body. This task required the students to draw and explain the blood-path as they knew it, and the tutors to draw and explain what they thought the students knew about the blood path. The explanations given while drawing were audiotaped and transcribed.

Results

Assessing and Diagnosing Beliefs

The students' and tutors' explanations and blood path drawings after sentences 24 and 76 were coded at both the proposition and the mental model levels. Due to a technical problem (failure of the audio taping machine), we were only able to recover the explanation data from 5 tutor-tutee pairs. Thus, only the explanation data of these 5 pairs will be used to analyze the

beliefs, whereas the drawing data from all 11 pairs (along with the accompanying explanations for the 5 pairs) were used for the mental model analysis.

Assessing template-based beliefs. A template of 38 beliefs relevant to the blood path as expressed in the text used in this study (Towle, 1989) was created. We can consider this to be the target normative knowledge that students are expected to learn about the blood path, given this text passage. A belief is a proposition size statement, such as *The heart is divided into 4 chambers*, or *The aorta is an artery.* Based on this target template, the explanation data generated after sentences 24 and 76 were coded in three ways. First, the explanation data were coded for the correct mentioning of each of these beliefs. Averaging across the data from sentences 24 and 76, tutors assumed that the students knew, on average, 21.5 out of the 38 beliefs from the template (Table 2). Students in fact knew an average of 18.9 beliefs. This means that tutors over-estimated students with knowing an additional 2.6 beliefs. Of the 18.9 correct student beliefs, tutors accurately detected on average 14.9 of these. This means that the tutors missed detecting 4 out of the 18.9 beliefs that students did know. Thus, tutors' accuracy at assessing what students knew was 78.8% (14.9/18.9).

Table 2.
Number of template-based beliefs (out of 38)

	STUDENTS KNEW	STUDENTS MISSED	TOTAL
TUTORS ASSUMED STUDENTS KNEW	14.9	6.6	21.5
TUTORS ASSUMED STUDENTS MISSED	4	12.5	16.5 (by "default")
Total	18.9	19.1	38
Tutors' assessment accuracy	78.8%	65.4%	
Overall assessment accuracy			72%

NOTE: Data based on 5 tutor-tutee pairs. The data cannot be subjected to a signal detection analysis because the numbers in the cells are not independent.

Conversely, "by default", tutors on average assumed that students missed 16.5 beliefs. "By default" means that the tutors did not explicitly express that students did not know these 16.5 beliefs (the remaining beliefs in our template), but we credited tutors with making that assumption. Students actually missed (either did not mention or mentioned incorrectly) on average 19.1 template beliefs (Table 2, column 3). This means that tutors under-estimated what students missed by 2.6 beliefs. Of these 19.1 missed student beliefs, tutors detected that students missed 12.5 of them. Thus, tutors' accuracy at assessing what target knowledge students missed

from the template was 65.4% (12.5/19.1). Overall, on average, tutors' *assessment* accuracy, based on the text template, was 72%, which seems fairly accurate[2].

Diagnosing alternative beliefs. Table 2 essentially shows how many template-based target beliefs students did know, and how many target beliefs students missed (did not mention or mentioned incorrectly). What about alternative beliefs? These would be beliefs that are not on the template, thereby not addressed by this specific text. These alternative beliefs can be acquired either from everyday experiences or inferred (correctly or incorrectly) from the text passage. Thus, although alternative beliefs have traditionally been conceived of as *false*, they can, however, be *true* in the normative sense (i.e., from the perspective of scientific knowledge at large). For example, in this particular passage, no mention was made of the relative size of the left atrium as compared to the right atrium, but students can sometimes infer correctly from the text that the left atrium is larger than the right atrium, and this would be a *true* alternative belief. Thus, the same protocols were further coded for the presence of alternative beliefs that were not on the template but the students knew, and that the tutors thought the students knew.

Table 3 shows that tutors assumed students knew, on average, 4.2 true alternative beliefs, but students in fact knew only 2.4 true alternative beliefs. So tutors over-estimated students with knowing more correct normative (or true alternative) beliefs than they actually did. Of the 2.4 true alternative beliefs that students did know, tutors accurately detected on average only one of them. Thus, tutors' diagnosis accuracy for true alternative beliefs was 42%. Tutors did not think that students had any false alternative beliefs, but in fact students had an average of 2.6 false

[2] Note that a signal detection analysis is the conventional method used to determine tutors' monitoring accuracy. However, it is unsuitable for use here because the numbers in the cells are not independent, since we did not actually know how many beliefs tutors thought students missed. We only assumed by default.

beliefs. Thus tutors under-estimated students with fewer false beliefs than they actually had. So, tutors' diagnosis accuracy for false beliefs was 0%.

Table 3.
Number of alternative beliefs

	TRUE ALTERNATIVE	FALSE ALTERNATIVE	
TUTORS ASSUMED STUDENTS KNEW	4.2	0	
STUDENTS KNEW	2.4	2.6	
Detection Accuracy	1.0	0	
Tutors' diagnosis accuracy	42%	0%	
Overall diagnosis accuracy			21%

NOTE: Data based on 5 tutor-tutee pairs.

Because alternative beliefs (whether true or false) are not stated in the text of this study, the tutors' accuracy in detecting them provides a very stringent test of their propensity (and perhaps ability) to diagnose students' alternative understanding. If tutors were accurately diagnosing students' alternative understanding, and not merely assessing what aspects of the text template students were missing, then tutors ought to be able to detect what alternative beliefs (either true or false ones) that students might have.

In sum, two results emerged. First, comparing Table 2 with Table 3, tutors were clearly not as aware of what students knew beyond the concepts explicitly stated in the text. Their

diagnosis rate for detecting the number of true alternative beliefs was only 42% compared with an assessment rate of 78.8% (Table 2) for detecting the number of correct template-based beliefs. Table 3 further shows that tutors' diagnosis rate on false alternative beliefs was 0%, compared to an assessment rate of 65.4% (Table 2) for detecting what students missed. Overall, tutors were clearly better at assessment (average accuracy of 72%) than diagnoses (average accuracy of 21%).

The second result that we did not predict was that tutors tended to overestimate students' correct template-based (Table 2) and true alternative knowledge (Table 3), and under-estimate students' incorrect template-based (Table 2) and false alternative knowledge (Table 3). Although these differences are not statistically significant (because of small n), the pattern of over- and under-estimation suggests that tutors assessed students' understanding from their own (presumably) normative perspective, so that they tended to make attributions of what students knew based on what they themselves knew, resulting in attributing students with knowing more correct and true knowledge and less incorrect and false knowledge. This pattern of attribution is further supported by tutors' willingness to attribute students with knowing true alternative beliefs (4.2 in Table 3) but reluctant to attribute students with knowing any false alternative belief (0 belief, see Table 3). In sum, although the data available for these analyses are meager, taken together, these patterns of results are all consistent with the interpretation that tutors tend to assess students' understanding from a normative, tutor-centered perspective, and not from an alternative, student-centered, perspective.

Assessing and Diagnosing Mental Models

Based on our prior findings (Chi et al., 1994), we assumed that codings based on individual beliefs tap shallow understanding; therefore it is important to evaluate tutors' diagnostic skills for deep understanding. Accordingly, we analyzed such deep understanding by determining what mental models the students were building to represent their understanding of the circulatory system, as well as what mental models the tutors thought students were building. Thus, this constitutes an analysis of the students' deep understanding at a different grain size.

Mental models can be captured by coding students' explanations of the components of circulation and their connections, as described in Chi et al. (1994). We have developed criteria to determine what constitutes a "single loop" model, what constitutes a "single loop with lungs" model, for a total of 7 mental models (see Figure 1 of Chi et al, 2001, and Figure 1 here shows 2 of the 7 models). Of the 7 mental models, 6 of them are flawed, as determined from prior work that captured the various degrees to which they are flawed.

Using all the drawing data from the 11 tutor-student pairs after sentences 24 and 76 (along with the transcribed audio taped protocols that were available for 5 of the 11 pairs), we captured the students' and the tutors' conceptions of their students' mental models in terms of the 7 mental models. Table 4 reports the results of the accuracy of tutor's mental model detection. The first column shows the frequency (14 out of 22, or 63.6%) students depicted the correct Double Loop mental model. (There were 11 students, but each was asked to draw twice, for a total of 22 samplings.) In contrast, the tutors assumed that students would depict the correct Double Loop model in 19 out of 22 cases (86.4%). Thus, tutors tended to over-estimate how often students would know the correct model. The tutors accurately detected that the students held the correct Double Loop model in 12 out of 14 cases. Thus, the tutors' assessment accuracy

(determined by the match between the individual student's model and the tutor's model of the student's model) was 86% (12/14).

The second column shows the frequency (8 out of 22) that students held one of the six flawed mental models. However, the tutors assumed that the students held a flawed mental model in 3 out of 8 cases, so that the tutors under-estimated the frequency that students held a flawed model. However, of the three occasions that the tutors assumed the students held flawed mental models, only one of them was in fact flawed. Moreover, the detection was inaccurate in terms of which flawed model it was. (The tutor thought the student held a No Loop model, but the student in fact held a Multiple Loop model.) Thus, the tutors failed completely to detect which flawed blood path model their students held; therefore their diagnosis accuracy was 0%.

Table 4.
Number of mental models

		STUDENTS KNEW		
		Correct (Double Loop)	Flawed (1 of 6)	Total
TUTORS ASSUMED	Correct	12	7	19
	Flawed	2	1	3
	Total	14	8	22
		(Assessment accuracy: 12/14=86%)	(Diagnostic accuracy: 0/8=0%)	

NOTE: Data based on 11 pairs, 22 instances at Sentences 24 and 76. We cannot report the result of this table in terms of hits and misses in a signal detection way because correctness and flawness are not binary constructs, since a flawed mental model can be one of six.

Overall, the analyses carried out at the mental model level mirror the same two patterns of results from the analyses undertaken at the proposition level. First, the tutors were significantly more accurate at assessing what correct (Double Loop) mental modes students held, than in diagnosing what flawed mental models students held ($\chi^2(1, N = 22) = 11.29, p < .001$). Second, the tutors tended to over-estimate the frequency of correct mental models and under-estimate the frequency of flawed mental models that students held, suggesting that they were assessing from their own normative perspective, and not diagnosing from the students' alternative perspective.

Thus, analyses at both the belief and the mental model levels show that tutors tended to over-estimate what students knew and under-estimate what students missed or knew in a flawed way. This overestimation appears to result from tutors' bias to expect that students had some normative knowledge even though students did not overtly express it. This bias worked against them when they had to detect students' alternative understanding (such as false beliefs and flawed mental models). This bias suggests that tutors monitored their students' understanding based on an assessment of what they expected the students to know, and not from an accurate diagnosis of what students in fact did or did not know.

Assessing Knowledge Deficits

The results of Tables 2-4 show that tutors were not particularly competent at diagnosing students' alternative understanding (in terms of false beliefs and flawed mental models). Instead, tutors seemed to be more competent at assessing students' normative understanding (in terms of the template-based beliefs and the Double Loop model). However, because the assessment

accuracies could have been biased from tutors' guessing of what correct knowledge students knew (on the basis of what they themselves knew) during the drawing and explaining task inserted after tutoring sentences 24 and 76, we can eliminate this bias and determine more authentically the extent of tutors' assessment competence by seeing how well tutors were able to detect *incorrect* knowledge *during* the tutoring sessions. That is, the prior analyses captured tutors' assessment accuracies by coding what tutors expressed in their drawings and explanations after sentences 24 and 76. The purpose of this analysis, to be described below, is to determine tutors' assessment capability embedded in the context of tutoring.

The analysis is based on a new coding of the entire tutoring protocols of all 11 tutor-tutee pairs. Each pair generated about two hours of tutoring protocols, and they were all transcribed. In these transcribed protocols, we defined knowledge deficits (borrowing a term from Graesser, et. al 1995) to be incorrect beliefs, expressed by the students, of one or more of the following kinds:

1) contradictory beliefs (as defined earlier);

2) incomplete but partially correct beliefs, such as stating "A valve" when a more complete answer should be "A semilunar valve";

3) vague beliefs, such as, in response to the tutor's statement "So, the pulmonary circulation is the only one that is from heart to lungs, lungs to heart?" and the student responded vaguely with "Like the bus?";

4) missing beliefs, in that the student failed to produce the requested information;

5) false beliefs (as defined earlier).

Knowledge deficits do not include trivial errors, such as mispronouncing words.

A total of 323 knowledge deficits were identified across 11 tutor-tutee pairs. These 323 knowledge deficits can be further differentiated as either tutor-initiated (i.e., expressed either as a result of answering the tutors' questions or explaining in response to tutors' scaffoldings) or student-initiated (i.e., expressed either while students were asking the tutors questions, or while self-explaining which was unprompted by the tutors). The first episode below shows a repeated knowledge deficit #73 (underlined; square brackets indicate our coding) expressed as a response to the tutor's query, whereas the second episode shows two distinct self-initiated knowledge deficits #21 and #22:

KD#73

T: Where does the oxygenated blood go after it enters the left atrium?

S: Through the whole body…. [Tutor-initiated]

T: What?

S: Whole body. [Tutor-initiated]

T: No. It goes to the left ventricle. [Tutor addressed]

KD#21, #22

T: Once equilibrium is established, the random movement of molecules continues, and equilibrium is maintained.

T: So

S: So the blood moves and it always tries to make equilibrium with every. [Student-initiated]

T: Right. With the, this is the, in the capillaries though. That's where the exchange is happening. [Not addressed] [Accepted]

S: So after one diffusion goes and they used it up, another diffusion,
So it just keeps on going and going. O.K. [Student-initiated]

T: Right. So, basically they're saying here that. The molecules that
they're talking about here are these oxygen- [Not addressed]
[Accepted]

Because we assume that tutors can detect knowledge deficits when they were expecting them by direct questions and other forms of elicitations, we further analyzed only the student-initiated knowledge deficits. Of the 323 knowledge deficits, 191 of them (about 59%) were student-initiated (i.e., they were generated while the students were self-explaining or asking the tutors questions, as shown in KD#21 and #22 in the preceding episode; they were not generated in the context of responding to what the tutors were eliciting, as in the case of KD#73 above). The critical question then is, how many of the student-initiated knowledge deficits were detected by the tutors? That is, how accurately can tutors monitor the presence of a knowledge deficit when they were not expecting one?

Addressing the Knowledge Deficits

One way to determine whether tutors detected students' knowledge deficits is to see whether or not they addressed them immediately. "Addressing" a knowledge deficit is operationalized to mean that the tutors were making the following kinds of content-relevant responses: giving direct feedback about the incorrectness of the knowledge deficit, scaffolding the student, or giving the correct answer. KD#73 embodied two of these attributes: The tutor responded to the knowledge deficit with a direct feedback "No," followed by the correct answer

"It goes to the left ventricle." "Not addressing" the knowledge deficits means that the tutors did not give a response that is relevant to the knowledge deficit. KD#21 and #22 shown above were student-initiated vague responses. In KD#22, it's not clear what the student meant by saying that "diffusion goes"; but in any case, the tutor did not address this vague comment, and proceeded to talk about molecules instead. The next two examples show student-initiated knowledge deficits that were addressed by the tutor.

KD#10

T: These are strong contractions because this blood is going to

S: Be pumped all through…

T: the entire body…in fact, these contractions are, they are massive, you know, You can feel your pulse in your wrist or whatever.

S: Every beat is another contraction? [Student-initiated]

T: In fact, the beats are a little different than your pulse. The pulse is basically the, the, the blood pressure from your arteries it is making the arteries expand and that is basically what you are feeling. [Tutor addressed]

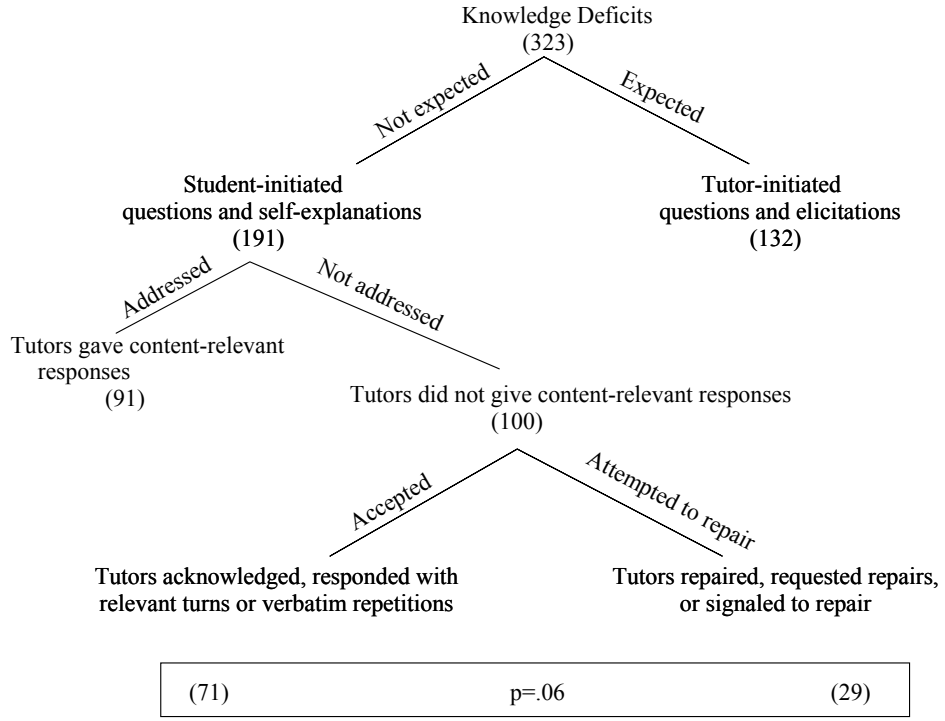
KD#33

T: Liver kind of filters out things, filters out the bad stuff that you eat.

S: When you smoke cigarettes, the liver- [Student-initiated]

T: No, actually when you smoke cigarettes, it's your lungs that do that. [Tutor addressed]

Using these coding criteria, the results showed that tutors addressed the student-initiated knowledge deficits slightly less than half of the time. That is, 91 out of the 191 student-initiated knowledge deficits were addressed (see Figure 3). This could mean that the tutors failed to detect 100 of these unexpected knowledge deficits (or 52%).



Establishing Common Ground

To be charitable to the tutors, one could interpret tutors' failure to address 52% of the unexpected knowledge deficits in two ways. One reason for not addressing them is the one provided above, namely that the tutors failed to detect them. On the other hand, since "addressing them" is defined in the context of explicitly giving direct feedback followed by *content-relevant* scaffolding or the correct answers, then an alternative interpretation of "not addressing" is that the tutors may simply fail to know *how to repair* the knowledge deficits, even

though they were detected. Tutors may not know how to repair the knowledge deficits if they either did not understand what the students expressed or the relationships between what the students were expressing and the normative knowledge.

To discriminate the subtle difference between whether tutors failed to detect knowledge deficits or failed to know how to repair them, we adapted coding methods from conversational analyses that can be interpreted without requiring explicit evidence of addressing, as defined above in terms of content-relevant feedback and scaffolding responses. That is, it is possible that a subtle conversational move may indicate detection even if no overt response, relevant in content to the knowledge deficits, was expressed. Accordingly, to test this possibility, we used Clark and Schaeffer's (1989) six common ground moves suitable to verbal protocols (excluding gestures and eye contact since we do not have this kind of data). We operationalized three kinds of moves (respond with the next relevant turn, acknowledge, display or demonstrate understanding through verbatim repetition) to be ACCEPT moves and three other kinds (repair, request repair, and signal to acknowledge) to be ATTEMPT to REPAIR moves. That is, ACCEPT moves would indicate that the tutors did not detect a knowledge deficit, whereas ATTEMPT to REPAIR moves might indicate that tutors have detected a knowledge deficit. For instance, take the following student-initiated knowledge deficit that has been originally coded as "not addressed" by the tutors, because the tutors did not give either explicit content-relevant feedback, or scaffolding, or any kind of answer. Instead, the tutor merely repeated what the student said.

<u>Tutor displays understanding by verbatim repetition:</u>

S: "How do you get poor muscle tone?"

T: "How would you get poor muscle tone?"

However, verbatim repetition, according to conversational analyses, is one type of ACCEPT move. On the other hand, the following student-initiated knowledge deficit was also not addressed by the tutor in the sense that the tutor neither scaffolded, gave feedback nor the correct answer. However, it can be coded as a request-to-repair move, in the context of conversational analyses:

Tutor requests repairs:

S: "O.K. Deoxygenated, it would be oxygenated blood."

T: "Deoxygenated or oxygenated?" ACCEPT moves can thus be interpreted to mean that the tutors failed to detect a knowledge deficit. On the other hand, ATTEMPT to REPAIR moves imply that the tutors might have detected a knowledge deficit, and attempted to repair. The analyses showed that there were far more total ACCEPT moves (71) than REPAIR moves (29) (or means of 6.5 versus 2.6 per tutor-tutee pairs, respectively, and the difference in means is marginally significant, $t(20)=1.92$, $p=.06$, see Figure 3). This suggests that, even using a more sensitive conversational type of analyses that credits tutors with detection when knowledge deficits were not addressed, tutors failed to detect 71 of the 191 unexpected knowledge deficits. So one can safely conclude that at minimum, 71 of them (or 37%) were truly not detected.

In sum, one might conclude that of the 191 student-initiated knowledge deficits, tutors detected at best 120 of them or 63% (that is, of the 191 student-initiated knowledge deficits, 91 of them were addressed and 29 more had evidence of attempts to repair). This suggests that tutors' assessment accuracy for *incorrect* beliefs (63%), was far less than tutors' assessment accuracy for template-based *correct* beliefs (of 78.8%) and for the correct mental models (86%). This supports our previous interpretation that the tutors' assessment accuracy was inflated by tutors' bias in guessing what correct knowledge students knew. Such bias suggests that the tutors

were attributing the students with knowledge that they themselves knew, from their own perspectives.

Although one can always interpret these results by assuming that tutors intentionally chose not to address a knowledge deficit or attempt to repair it, either because they thought that the knowledge deficit was trivial, or that it was not relevant, or that it was not the right moment to repair it, we cannot dismiss our evidence by surmising what tutors might be thinking. We can only interpret what overt actions tutors took.

Discussion

The analyses reported above, derived from coding at the proposition and the mental model levels (shown in Tables 2-4) gave a systematic picture of how novice tutors monitored students' understanding. Several conclusions arise from these analyses. First, re-confirming the suggestive evidence in the literature, there is no doubt that everyday tutors failed to accurately *diagnose* students' alternative (false and flawed) knowledge, either at the proposition level or at the mental model level. Second, at first glance, tutors seemed to be extremely competent at *assessing* what text-based (or normative) knowledge students had (again, at both the belief and the mental model levels). However, this accuracy in assessing what normative knowledge students had seemed to result from tutors' general bias to over-attribute students with normative knowledge and under-attribute them with alternative knowledge. This interpretation was further supported by looking at what *incorrect* knowledge (or knowledge deficits) tutors can accurately detect, while tutoring. Using a more sensitive conversational analysis, we found that tutors were not detecting over a third (37%) of the student-initiated knowledge deficits.

Overall, the results reported in this paper, using more direct and sensitive analyses with a larger sample of novice tutors, unambiguously support the conjecture in the literature that tutors (whether experienced or not) do not accurately *diagnose* students' alternative understanding; they are far better at *assessing* the extent students' understanding deviates from the normative. But even in assessment, they inflated their judgment toward assuming that students had more complete understanding than they actually did. This suggests that tutors monitored students' understanding from the tutors' perspective and not from the students' perspective [3].

The interpretation that tutors were operating with a normative model and students were often operating with their own idiosyncratic flawed models also provides evidence to address the notion whether convergence between the tutors and the students had occurred. That is, the literature on collaborative learning suggests that collaborating dyads often converge toward shared meanings (Rochelle, 1992), in the sense of establishing common ground. There is little evidence of it here in a tutoring context. Analyses of moves that seek common ground show that when tutors failed to address students' knowledge deficits, this failure is more likely due to a

[3] Our domain of inquiry was a conceptual one, so in what way can we generalize our results to a procedural domain, such as tutoring problem solving skills? That is, do tutors diagnose in procedural domains? Aside from the brief literature that alludes to this question that was reviewed earlier, this question can be answered more directly by examining the kind of monitoring that intelligent tutoring systems (ITS) have implemented. There are basically two kinds of intelligent tutoring systems. One kind uses an overlay model and the other kind uses a bug-based model. An overlay model simply determines whether an action taken by the student is correct or incorrect, with respect to the correct solution that the system embodies. For example, suppose at a specific point, the ITS expects an equation such as F=ma, but the student wrote $F=mv^2$, the ITS will simply recognize that the equation is an incorrect one. The monitoring embodied in this type of ITS is analogous to template-based assessment. A second kind of ITS is a bug-based model. In this kind of model, the system represents the alternative procedures (or buggy rules) that the student might use, on the basis of prior analyses (Brown & VanLehn, 1980). Such a system is capable of diagnosing on-line students' use of *alternative* procedures, but the set of alternative procedures embedded in the ITS must be identified a priori. In subtraction, for example, Brown and VanLehn (1980) analyzed a large corpus of students' subtraction errors, and captured through such analyses some of the students' bugs, such as taking the smaller number from the larger number in the same column, regardless of whether the larger number is at the top or at the bottom. Such procedural bugs can then be represented in the ITS, which can then be used to successfully diagnose students' alternative procedures. Thus, ITS do implement a diagnostic capability that apparently is not privileged to human tutors, but the definition of diagnosis is similar. However, a lack of correspondence between our meaning of *diagnosis* and the ITS may be at the level of the mental model. It is not clear what diagnosing a flawed mental model corresponds to in a procedural domain.

failure of detection and not a failure in *not* knowing how to repair a knowledge deficit. Thus, the processes of tutoring and peer collaborating may be dramatically different.

In order to address the question raised earlier of whether or not tutors are adaptive, by extrapolation, one would have to conclude that because tutors were not very accurate at monitoring students' alternative understanding, they could not have been adaptive in their pedagogical skills, such as customizing their feedback and explanations based on the students' alternative understanding. Instead, their explanations (see the data in Chi et al. 2001, for example) must have been delivered very much in the spirit of knowledge displays and knowledge telling, and their feedback must have been based largely on an assessment of the correctness of students' responses in the context of normative knowledge. Moreover, because tutors were so inaccurate at diagnosing students' flawed mental models, they could not have been customizing their feedback and explanations at a global mental model level. Instead, feedback must have been given at the local level, which tends to engender shallow understanding.

Although we began the paper favoring the hypothesis that tutors can accurately assess what students know based on overlaying the normative template over what the students knew, and rejecting the hypothesis that tutors can diagnose what alternative knowledge students knew, our data show that tutors were less than adequate even at assessment. Not only were tutors totally unable to predict what alternative knowledge students might know (i.e., they were downright dismal in diagnosing students' false beliefs and alternative flawed mental models), but tutors appeared to be underwhelming even in their ability to assess what incorrect knowledge (knowledge deficits) students had (around 63% at best). Therefore, if novice tutors cannot accurately monitor students' normative understanding and alternative understanding, we cannot

attribute the overall tutoring effectiveness largely to the tutors' pedagogical skills, since tutoring effectiveness is achieved even with everyday inexperienced tutors. Thus, the results of these analyses indirectly support the hypotheses we raised earlier (Chi, et al., 2001), that tutoring effectiveness may arise from the students' constructive learning while being tutored, as well as some forms of (yet to-be-determined) interactions, and not necessarily from tutors' adaptiveness in the context defined here[4].

[4] Tutors may be adaptive in a different context, such as in post-solution discussions (Katz, Allbritton & Connelly, 2003).

Bibliography

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.

Brown, J.S., & VanLehn, K. (1980). Repair theory: A generative theory of bugs in procedural skills. *Cognitive Science*, 4, 379-426.

Chi, M.T.H. (in press). Common sense conceptions of emergent processes. *Journal of the Learning Sciences.*

Chi, M.T.H. (2000). Self-explaining: The dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in Instructional Psychology* (pp. 161-238). Mahwah, NJ: Lawrence Erlbaum Associates.

Chi, M. T. H. (1996). Constructing self-explanations and scaffolded explanations in tutoring. *Applied Cognitive Psychology*, 10, S33-S49.

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., & Roscoe, R.D. (2002). The processes and challenges of conceptual change. In M. Limon and L. Mason (Eds.), *Reconsidering Conceptual Change: Issues in Theory and Practice* (pp. 3-27). Kluwer Academic Publishers, The Netherlands.

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

Cho, B.I., Michael, J. A., Rovick, A. A., & Evens, M. W. (2000). An analysis of multiple tutoring protocols. In: G. Gauthier & C. Frasson & K. VanLehn (Eds.), Intelligent Tutoring Systems: 5th International Conference (pp. 212-221). Berlin: Springer.

Clark, H., & Schaefer, E.F. (1989). Contributing to discourse. *Cognitive Science*, 13, 259-294.

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

Collins, A., & Stevens, A. (1982). Goals and methods for inquiry teachers. In R. Glaser (Ed.), *Advances in Instructional Psychology, Vol. 2 (pp. 65-119)*.Hillsdale, NJ: Lawrence Erlbaum Associates.

DeLeeuw, N. (1993). Students' beliefs about the circulatory system: Are misconceptions universal? In Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society, (pp. 389-393). Hillsdale, NJ: Lawrence Erlbaum Associates.

Fitz-Gibbon, C.T. (1977). *An analysis of the literature of cross-age tutoring.* Washington, D.C.: National Institute of Education (ERIC Document Reproduction Service No. ED 148 807)

Gentner, D. & Stevens, A.L. (1983). Mental Models. Hillsdale: Lawrence Erlbaum Associates.

Graesser, A. C., Person, N., & Magliano, J. (1995). Collaborative dialog patterns in naturalistic one-on-one tutoring. *Applied Cognitive Psychology*, 9, 359-387.

Jeong, H. (1998). Knowledge co-construction during collaborative learning. Unpublished Ph.D. thesis, University of Pittsburgh.

Katz, S., Allbritton, D. & Connelly, J. (2003). Going beyond the problem given: How human tutors use post-solution discussions to support transfer. *International Journal of Artificial Intelligence in Education*, 13, 79-116.

McArthur, D., Stasz, C., & Zmuidzinas, M. (1990). Tutoring techniques in algebra. *Cognition and Instruction*, 7(3), 197-244.

Merrill, D.C., Reiser, B.J., Ranney, M., & Trafton, J.G. (1992). Effective tutoring techniques: a comparison of human tutors and intelligent tutoring systems. *The Journal of the Learning Sciences*, 2(3), 277-306.

Putnam, R. T. (1987). Structuring and adjusting content for students: A study of live and simulated tutoring of addition. *American Educational Research Journal*, 24(1), 13-48.

Rochelle, J. (1992). Learning by collaboration: Convergent conceptual change. *Journal of the Learning Sciences*, 3 & 4, 235-276.

Stevens, A.L., & Collins, A. (1977). The goal structure of a Socratic tutor. Proceedings of Association for Computing Machinery National Conference, Seattle, Washington.

Towle, A. (1989). *Modern Biology*. New York: Holt, Rinehart and Winston.

Vosniadou, S. & Brewer, W.F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology*, 24, 535-585.