Running head: OBSERVING TUTORIAL DIALOGUES

Observing Tutorial Dialogues Collaboratively:
Insights about Human Tutoring Effectiveness from Vicarious Learning

Michelene T. H. Chi
Department of Psychology
Learning Research and Development Center
Phone: 412-624-7456
Fax: 412-624-7904
Email: chi@pitt.edu

Marguerite Roy
Learning Research and Development Center
Phone: 412-624-1343
Fax: 412-624-7904
Email: mar982@pitt.edu

Robert G.M. Hausmann
Learning Research and Development Center
Phone: 412-624-7536
Fax: 412-624-7904
Email: bobhaus@pitt.edu

Keywords: tutoring; human tutorial dialogue; physics; problem solving; learning; learning from observing; vicarious learning

Abstract

The goals of this study are to evaluate a relatively novel learning environment, as well as to seek greater understanding of why human tutoring is so effective. This alternative learning environment consists of pairs of students collaboratively observing a videotape of another student being tutored. Comparing this collaboratively observing environment to four other instructional methods – one-on-one human tutoring, observing tutoring individually, collaborating without observing, and studying alone – the results showed that students learned to solve physics problems just as effectively from observing tutoring collaboratively as the tutees who were being tutored individually. We explain the effectiveness of this learning environment by postulating that such a situation encourages learners to become active and constructive observers through interactions with a peer. In essence, collaboratively observing combines the benefit of tutoring with the benefit of collaborating. The learning outcomes of the tutees and the collaborative observers, along with the tutoring dialogues, were used to further evaluate three hypotheses explaining why human tutoring is an effective learning method. Detailed analyses of the protocols at several grain sizes suggest that tutoring is effective when tutees are independently or jointly constructing knowledge with the tutor, but not when the tutor independently conveys knowledge.

Observing Tutorial Dialogues Collaboratively:

Insights about Human Tutoring Effectiveness from Vicarious Learning

1. Introduction

Although learning of subject matter domains can occur with conventional methods of instruction that are based on a transmission view of learning, much recent work has focused on instruction based on a more constructivist view of learning. The two most widely studied and implemented methods are human tutoring and collaborating with a peer. Compared to learning in a traditional classroom, learning from face-to-face human tutoring is extremely beneficial and has the largest effect size, ranging from 0.4 to 2.0 (Bloom, 1984; Cohen, Kulik & Kulik, 1982); followed by learning from collaborating with a peer, with a substantially smaller effect size, ranging from 0.21 to 0.88 (Johnson & Johnson, 1992; Slavin, 1990). Studies that have directly compared interacting with a tutor (or experimenter) with collaborating in pairs have also found tutoring to be more effective than collaborating (Pilkington & Parker-Jones, 1996). Besides these two widely studied methods, many other creative methods of learning are being explored, such as learning by observing (McKendree, Stenning, Mayes, Lee, & Cox, 1998) and learning by teaching an animated agent (Biswas, Schwartz, Bransford, & TAG-V, 2001; Schwartz, Blair, Biswas, Leelawong, & Davis, in press).

Although human tutoring is extremely effective, it is not cost-effective to scale up. Current usage of human tutoring in schools tends to be pull-out systems that are available to only a few students. Developing intelligent tutoring systems is also costly and difficult to implement, although substantial progress is being made (e.g., by the Pittsburgh Science of Learning Center, see www.learnlab.org).

The research reported in this paper has both a theoretical and a pragmatic goal. The theoretical goal is to gain further understanding of why face-to-face human tutoring is so effective. Better understanding of how tutoring is effective will help improve designs for intelligent tutoring systems (ITSs) as well as prescribe designs for alternative learning environments that might be more easily developed, implemented, and scaled up. The pragmatic goal of this research is to test such an alternative learning environment, one that leverages the advantages of learning from tutoring and collaborating, with learning from observing, into a single learning environment. Very little research has explored the effectiveness of learning from observing another student learn, and moreover, the results have been inconsistent. We propose an explanation that can not only account for the discrepant results in the literature, but can also serve as a method to optimize the effectiveness of learning from observing. We refer to this potential explanation as the *active/constructive/interactive observing hypothesis*, and it refers to how actively engaged and constructive the observers are. This hypothesis is derived more generally from our earlier finding that being constructive facilitates learning, the general self-explaining effect (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Chi, de Leeuw, Chiu, & Lavancher, 1994).

This paper begins by analyzing the tasks and reinterpreting the results from the learning from observing literature. Then the paper introduces our *active/constructive/interactive observing hypothesis* and describes our pragmatic goal, which is to test the effectiveness of an alternative learning environment by comparing it with four other learning methods. We then explore our theoretical goal by elaborating three tutoring hypotheses and seek support for these hypotheses with in-depth analyses of

the tutoring protocols. In order to gain further understanding of tutoring effectiveness, we examine learning from both the perspective of the tutees as well as the perspective of the observers. The paper concludes with a discussion of the conditions under which learning from tutoring and learning by observing tutoring can be successful and ways in which tutoring can be usefully expanded and scaled up.

2. Learning from Observing

There is currently somewhat of a discrepancy in the literature with respect to the benefit of learning from observing. We start by defining how observational learning has been investigated, then address the discrepancy issue.

We use the very neutral terms *learning from observing* or *learning vicariously* to encompass learning contexts that include both *watching* someone else learn (*watching* refers to observing visual inputs) and *overhearing* the ensuing dialogues between a learner and an instructor (*overhearing* refers to observing auditory inputs). Aside from the distinction between *watching* and *overhearing*, another important distinction that needs to be made is between learning from observing someone else *learn* and learning from observing someone else *act or behave*.

Learning from observing *actions* or *behavior* has been explored fairly extensively in at least three or four different areas of research. In social psychology, for example, studies have been undertaken to see how one learns from watching someone act aggressively (Bandura, 1969; 1986). In these cases, it is not necessary for the persons we watched to have exhibited learning (i.e., they were not learning to act less aggressively); they merely have exhibited some actions from which we the observers might learn. In short, they are the actors and we are watching them model some behavior. The role of

discourse is often not explicitly addressed in such cases because much of this work focuses on physical skills (thus, the *overhearing* component of observing is absent). By and large, highly effective learning of physical skills does take place by *watching* overt behavior of individual actors, even without exposure to discourse (i.e., without *overhearing*).

Learning from observing modeling has also been studied in work settings (Latham & Saari, 1979) and on-the-job apprenticeship learning. More recently, learning from observing actions has been studied with renewed interests by neuroscientists and developmental psychologists, termed *imitative learning*. Imitative learning is an ability that is uniquely human and not available to primates, in the form that requires the observer, such as a child, to transform and take on the perspectives of others (Meltzoff, 2005). Imitative learning is assumed to play an important role underlying the preservation of cultural practices (Meltzoff, 2005; Tomasello, 1999).

It is not totally surprising that one can learn from observing physical skills, since the actions underlying physical skills can be overtly modeled, in a more-or-less one-to-one correspondence. For example, knitting consists of making loops that can be modeled in a step-by-step way in terms of the direction and sequencing of the knitting needles (although the pattern or the design must be planned ahead), so one would predict that an observer can learn to knit by watching all the intervening steps.

But can one learn from observing mental or cognitive skills? Cognitive skills can include both *learning skills* (such as self-explaining, asking questions) and other *task-specific skills* (such as problem solving). A few recent studies have shown that one can also learn *learning skills* from observing. For example, students can learn to *ask*

questions by observing (watching and overhearing) an animated agent ask questions

(Craig, Gholson, Ventura, Graesser & the Tutoring Research Group, 2000), or learn to

collaborate by observing others collaborate (Rummel & Spada, 2005).

There are two important characteristics to note about learning these *learning skills*

from observing. One characteristic is that the observers were not expected to learn the

content of the articulations made by the actors. For example, the observers in the Rummel

and Spada (2005) study were not expected to learn the diagnosis of the panic disorder

case. Rather, the goal was to learn the *skill* of collaboration. A second characteristic is

that the actors themselves were not learning. For example, the animated agent in the

Craig et al. (2000) study was not learning. In short, in these two studies, the observers

were learning the *learning skills* of *asking questions* and *collaborating*, rather than

learning *the content* of the questions asked by the animated agent or the topic of the

collaborative diagnosis**.** Thus, the question remains as to whether one can learn the

content of cognitive skills (or *task-specific skills* such as problem solving) from

observing.

In Chi and Bjork (1991), we had argued that task-specific cognitive skills that

involve a many-to-one mapping between the processing steps and the observable overt

outputs might be much more difficult to learn from watching. For instance, when an

expert or teacher solves a math or physics problem on the board, if we *watch* only the

overt actions alone (such as writing equations), then the number of equations written does

not correspond to the complexity of the underlying reasoning steps. One solution to

overcoming the many-to-one correspondence between the covert reasoning steps and the

overt actions of solving complex problems is to have the expert externalize her thinking

or make her thinking visible by giving explanations in a monologue mode (Collins, Brown & Holum, 1991). This would add an *overhearing* component. However, making thinking visible does not overcome the dual problems that (a) experts are notorious for not being able to express all their thinking because much of their knowledge is tacit (Nisbett & Wilson, 1977), and (b) even if the experts could express their thinking elaborately as in giving monologue didactic-like explanations, it is becoming increasingly clear that students do not learn very well from hearing such explanations (Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001; VanLehn, Siler, Murray, Yamauchi & Baggett, 2003). For example, we already know that listening to an instructor, an expert, or a peer giving monologue explanations is less facilitating to learning when compared to many other ways, such as listening to dialogues (Driscoll, Craig, Gholson, Ventura, Hu, & Graesser, 2003; Fox Tree, 1999), being scaffolded (Chi et al., 2001), having the students self-generate explanations (Chi, Bassok, Lewis, Reimann & Glaser, 1989; Webb, 1989) or self-generate questions (King, 1992).

To conclude, however, that complex cognitive tasks may not be learnable from observation, even if experts make their thinking visible in an explanatory monologue mode, may have been premature because it is possible that one can learn from observing if the observers can *overhear* not the monologue, but the dialogue between a learner and an actor/instructor. For example, instead of watching just one actor explaining and modeling a task-specific cognitive skill (such as a teacher explaining and solving a problem out-loud), can observers learn by *watching* and *overhearing* the dialogue between a teacher and a student? In fact, this would be a situation in which an observer is observing someone else *learn*. McKendree et al. (1998), and Stenning, McKendree, Lee,

Cox, Dineen and Mayes (1999) assume that access to dialogue is a critical component to learning vicariously. Two conditions in the Cox, McKendree, Tobin, Lee and Mayes' (1999) study test this assumption. They compared learning materials created by tutor-tutee dialogue (in which a novice student constructed tree diagrams with the help of an expert), captured as a movie and shown to observers, with learning materials consisting of only animation of the tree diagramming. Observers performed better when they watched a movie and overheard tutor-tutee dialogue than watching only an animation of diagramming, in which there was no verbal communication. Thus an observer could learn to diagram from *watching* and *overhearing* more so than just from *watching,* suggesting that *overhearing* is a critical component of learning from observing.

However, an earlier study by Schober and Clark (1989) seems to suggest the opposite conclusion. In the Schober and Clark (1989) study, the overhearers heard recordings of dialogues between pairs of participants in which one participant (analogous to a tutor*)* described the sequencing of 16 Tangram figures and the other participant (analogous to a tutee) had to order them as instructed. Each overhearer's task was also to sequence the Tangrams. The result was that the overhearers could not sequence them as accurately as the tutees who could converse with the tutor. Schober and Clark's conclusion was that understanding can only be built up by *participating* in dialogues, in which the participants could jointly build their common ground. Accessing dialogues from *overhearing* was argued to be not as effective as *participating* in dialogues.

At first glance, Cox et al.'s (1999) results seem to contradict the findings of Schober and Clark (1989) with respect to the benefit of overhearing dialogues in vicarious learning. However, the two studies are actually incomparable. In the Cox et al.

(1999) study, they were comparing two conditions of observing: *watching plus overhearing* versus just *watching*. Not surprisingly, *watching plus overhearing* is better than *watching* alone, as recommended by the "making thinking visible" approach to instruction (Collins et al., 1991). The Schober and Clark (1989) study, on the other hand, compared the conditions of *participating (*or interacting with a tutor*)* versus *observing* such interactions. Moreover, the former study measured learning whereas the latter study did not.

The Schober and Clark assumption, that only by participating in interactions can mutual understanding evolve, is also supported in the results of Craig, Driscoll, and Gholson (2004) and Craig, Sullins, Witherspoon and Gholson (2006). Across four experiments on learning, comparing interacting with a virtual tutor versus observing such interactions that were captured on video files (thus the observing included both *watching* and *overhearing*), they found that students interacting with a virtual tutor produced greater learning effects from pre- to post-test, than observing such interactions. In two of the experiments, the differences were significant, and in two other experiments, the differences were not significant, although in the same direction. Based on these preliminary findings, averaging across their four experiments gives us an effect size of 1.78 for interacting with a tutor and 1.12 for observing such interactions. In short, interacting with a tutor seems to be a more effective form of learning than observing such interactions.

In sum, judging from the difference in the relative magnitude of these effect sizes as well as the effect sizes reported in the literature on tutoring (ranging from $0.4 - 2.0$) and collaborating (from $0.21 - 0.88$), we can tentatively conclude that learning task-

specific cognitive skills from observing is possible when the observers have access to dialogues (with an estimated effect size of 1.12), although it appears to be less effective than learning from participating in tutoring, but perhaps more effective than learning from collaborating.

3. The Active/Constructive/Interactive Observing Hypothesis and
An Alternative Learning Environment

If learning from observing is in fact not as effective as learning from being tutored, is there a way to optimize learning from observing to make it comparable to the level of learning from tutoring? Essentially, the results in the literature suggest that interacting with a tutor is better than watching and overhearing. But is it interacting with a tutor that is important for learning or interaction *per se?* In all the studies cited above, the observers were observing individually with no opportunity to be interactive, so that they may or may not be actively engaged. Could the observers learn more if they had opportunities to be active or constructive or interactive? In this paper, these three terms will be used interchangeably and loosely. (More specific definitions, the underlying processes, and discriminations among the three terms—active, constructive, interactive—will be spelled out in a forthcoming paper.)

We tested our *active/constructive/interactive observing hypothesis* in a pilot study by encouraging individual overhearers to be actively engaged and constructive by requiring that they self-explain while *overhearing* an audiotaped recording of a tutor giving instructions to a tutee on how to put together a portion of an AM radio kit (Chi, McGregor & Hausmann, 2000). The design was essentially a replication of the Schober and Clark (1989) study but using a different task and requiring the overhearers to be

active. The overhearers' success in putting together the radio kit was compared to a tutee group who received the same instruction, but could interact directly with the tutor by asking questions. There were no significant differences in performance between the two groups. This null effect is consistent with our *active/constructive/interactive observing hypothesis*, suggesting that the difficulties sometimes encountered by overhearers may be attributed to their passive stance. Thus, the results in the literature showing the better performance of the participants who interacted with a tutor than participants who merely observed such interactions, may be explained by the lack of opportunities to actively interact *per se*, and not necessarily interacting with a tutor.

To further test our hypothesis, the present study implemented an alternative learning environment which involved giving the observers opportunities to interact, not with a tutor but with a peer. This environment consisted of having pairs of students collaboratively observe another student (a tutee) being tutored on how to solve a problem while simultaneously solving the same problem that is being tutored. It is essentially the same design as the Schober and Clark (1989) study, with the exceptions that pairs of students collaboratively observed while solving a problem, and learning gains were measured. This learning environment essentially capitalizes on the benefits of learning from tutoring and collaborating, in the context of learning from observing *someone else learn*. Learning in this target condition—*observing tutoring collaboratively*—is contrasted with learning from tutoring, collaborating, observing alone, and studying alone. We hope to provide direct evidence for the benefit or non-benefit of this alternative vicarious learning environment, thus testing the *active/constructive/interactive observing hypothesis*.

3.1   Method

3.1.1   Participants

One highly experienced teacher was selected to serve as a tutor in this study. This Tutor (henceforth capitalized to refer explicitly to him) possessed a Ph.D. degree in physics and had taught university-level physics for over 30 years. Moreover, this Tutor is a member of the Andes tutoring project (VanLehn, Lynch, Schulze, Shapiro, Shelby, Taylor, Treacy, Weinstein, & Wintersgill, 2005) in which he acted as the in-house domain expert. Therefore, he is somewhat familiar with the research issues, such as knowing the benefit of encouraging students to be constructive. We intentionally used a single tutor in order to examine the effect of tutee variability.

In addition to our Tutor, 70 undergraduates from the University of Pittsburgh participated as paid volunteers. Ten of these undergraduates participated in the Tutoring condition (henceforth referred to as the Tutees). The remaining 60 participants were assigned to four non-tutoring conditions. Each participant or pairs of participants in the non-tutoring conditions were yoked to one Tutee by gender so that an equal number of male and female participated in each condition. Unfortunately, due to this yoking procedure, complete random assignment to condition was not possible because we had to collect the tutoring data before collecting data from the remaining conditions. Furthermore, participation in the Tutoring condition required the Tutees to consent to allowing their videotapes to be shown to other participants in two of the conditions. Consequently, participants were assigned to the Tutoring condition before being randomly assigned to the remaining conditions.

All student participants were selected to have had taken at least one physics course at the high-school level, but no physics courses at the college-level. Prior to the study, each participant listed all the physics courses they had taken in high school and the grade(s) they received. There were no differences among conditions in either the number of courses taken or in their self-reported grades.

3.1.2   Materials

The problem-solving domain is quantitative kinematics, based on the materials in Chapter 5 of the classic physics textbook *Fundamentals of Physics* (Halliday and Resnick, 1981), dealing with the application of Newton's three laws of motion to problem situations (e.g., a block on an inclined plane, two blocks joined by a string supported by a pulley). Three kinds of problems were used in this study. The pre-test and post-test problems were used for assessment, and the tutoring problems were used in the intervention.

The *pre-test* consisted of four problems selected by the Tutor from the back of Chapter 5 that covered the major concepts and principles, such as weight force, normal force, compound body, tension, mass, acceleration, and Newton's three laws of motion. The easiest problem served as a warm-up problem. These four pre-test problems are shown in the first column of Appendix A. The appendix in its entirety can be found at http://www.cogsci.rip.edu/CSJarchive/Supplemental/Index.html.

For *tutoring*, the Tutor designed five problems that incorporated the same set of concepts and principles as were used in the pre-test. These tutoring problems (consisting of a warm-up problem and Problems 1-4), ordered by level of difficulty, with the Tutor's

rationale for constructing them, and their correspondences to the pre- and post-test problems, are shown in the middle column of Appendix A.

For the *post-test*, the experimenters designed four problems that roughly corresponded to Problems 1-3 of the pre-test and Problems 1-4 of the tutoring problems. (See Problems 1, 2, 3 & 4 in the post-test column of Appendix A). Problems 1-3 in the pre-test and post-test had identical problem situations (e.g. an inclined plane problem where the goal is to find tension), but they may differ in the values of the variables (i.e., for the angle, the forces, the masses), and some of the post-test problems have additional question parts. But essentially the deep structures of the problems were similar, in terms of what principles to apply. Problem 4 on the post-test did not correspond to any problem in the pre-test. Because Problem 4 was so difficult that no student solved any parts of it correctly, this post-test Problem 4 was eliminated from further analyses.

Besides the physics text, the assessment and tutoring problems, 10 videotapes were also created from the tutoring sessions of the Tutor with each of the 10 Tutees. Each videotape consisted of the Tutor tutoring a Tutee on the warm-up and two of the four tutoring problems. The average duration of the videotapes was 79.86 minutes.

3.1.2.1 Scoring of the post-test problems. The Tutor was asked to provide worked-out solution steps to the pre-test and post-test problems. A step is a written or spoken solution line, often an equation, reflecting a component of the problem-solving procedure, such as finding tension by applying Newton's Second Law. Each step can be differentiated as either *shallow* or *deep*. A *shallow* step reflected computation and the use of convention or notation, and a *deep* step required the application of physics concepts and principles.

Each step in solving the corresponding pre-test and post-test problem types was then compared. Some steps were repeated in both the pre- and the post-tests and these will be referred to as *matched,* and other steps were *unique* to solving either the pre-test or the post-test problems. A breakdown of the number of matched and unique steps can be found in Table 1. The majority of unique steps in the post-test were embedded in additional questions posed in parts b and c of a problem. There were a few unique steps in the pre-test because the pre-test were actual problems selected from the textbook whereas the post-test problems were constructed in such a way as to optimize assessment of learning from the tutored problems.

Insert Table 1 about here

3.1.2.2 Problem-solving models. The Tutor was also asked to explain how to solve each of the *tutoring* problems he constructed. His transcribed verbal protocol were used to create models of how to solve each of the tutoring problems in terms of problem-solving nodes, as shown in Figure 1 (for tutoring Problem 2). Each node corresponded to a state in a problem space. These models were later used to code and score the tutoring protocols.

Insert Fig. 1 about here

3.1.3   Design

The design for this study consisted of five conditions. The benchmark Tutoring condition, the target Observing Collaboratively condition, and three others conditions that served as various controls. In the Tutoring condition, each Tutee interacted individually with the same Tutor to solve three tutoring problems (the warm-up plus 2 of the 4 tutoring problems, chosen by the Tutor). Exactly which of the two tutoring problems were used with each Tutee varied across Tutees, as the selection was left to the Tutor's discretion. (Giving the Tutor a choice of problems was intended to mimic an authentic tutoring situation.) To ensure that the same set of problems was solved an equal number of times across all conditions, each participant or pair of participants in the other four conditions were yoked to one of the participants in the tutoring condition.

In the Observing Collaboratively condition, each pair of participants (henceforth referred to as Collaborative Observers) watched a videotape of one of the tutoring sessions and together they tried to solve the same three problems as were being solved by the Tutee in the tutoring tape. In the Collaborating condition, each pair of participants (henceforth referred to as Collaborators) solved one set of three problems (yoked to a Tutee set) collaboratively with no exposure to tutoring, but they were given access to the physics text from which they had studied Chapters 1-5. Thus, while this condition allowed for interaction with a peer, it controlled for the effects of observing tutoring. In the Observing Alone condition, each participant (Lone Observer) watched a videotape of one of the tutoring sessions individually and tried to solve the same problems as were being solved in the tutoring tape. Finally, in the Studying Alone condition, each participant (Solo Solver) tried to solve three problems (again yoked to a Tutee set) with

the text available as a resource. Table 2 summarizes the five conditions, the number of

participants in each condition, the learning resources available to each condition, and the

learning activities in which the participants could engage.

Insert Table 2 about here

3.1.4. Procedure

Following the procedure of our earlier work (Chi et al., 1989), the experiment

consisted of four distinct phases for all conditions: background (Chapters 1-4), Chapter 5

plus pretest, intervention, and post-test. During the first background phase, each

participant studied the first four chapters of Halliday and Resnick (1981), covering basic

background knowledge such as units of measurement, vectors, velocity as it relates to

motion in one dimension, concepts of force and mass, as well as how these concepts

relate to acceleration. Participants had to learn these first four chapters individually in our

lab to a criterion level, defined as attaining a score of at least 80% correct on a pre-

selected subset of problems and questions from the end of each chapter. If they failed to

achieve this standard, they were told which problems were wrong and were then required

to re-study the materials and re-solve the problems until they met the criterion. This

background phase took on average 6.34 hours to complete (ranging from 6.22 to 6.55),

occurring over several sessions. There were no significant overall or pair-wise differences

between conditions on this measure.

After attaining mastery of Chapters 1-4, the participants individually studied all of Chapter 5 in the second phase. Then they were asked to solve the pre-test problems, *with the text available to them*, in order to reduce stress and mimic a more authentic learning context. No feedback was given on the correctness of their solutions to the pre-test problems, nor were they required to reach any criterion. Thus, *the pre-test assessed how well students could learn Chapter 5 materials on their own*, having presumably mastered the prior (Chapters 1-4) knowledge to the same degree (80% criterion).

The first 10 participants that were recruited and had completed the pre-test were assigned to the Tutoring condition. Then the remaining 60 participants were randomly assigned to the other four treatment conditions for the third phase.

All problem solving took place on a large whiteboard so that we could capture on videotape an accurate record of all ongoing problem-solving activity. For all five conditions, the participants were audio- and videotaped. The Tutor was encouraged to keep the sessions to approximately one and one-half hours in length.

The Collaborative Observers in the Observing Collaboratively condition were given one of the tutoring tapes to watch on a large television along with the same three problems to solve. They were also required to perform their problem solving on a whiteboard with each member using a different colored marker. The Collaborative Observers were further told that they may discuss the videotape and their understanding of the materials at any time, and that they could stop, rewind, or fast forward any section of the tape. Allowing the participants to manipulate the tape in this way mimics potential authentic learning situations involving videotapes (such as a virtual or real classroom).

The Collaborators in the Collaborating condition were given a set of three problems to solve along with a copy of Chapters 1-5 from the Halliday and Resnick text. The Collaborators were encouraged to discuss and help each other solve the problems while working on a whiteboard.

The Lone Observers in the Observing Alone condition received similar instructions to those in the Observing Collaboratively condition. Rather than being encouraged to work with a partner (since they had no partners), the Lone Observers were told that they could talk out loud whenever it felt natural to do so, and they could stop, pause, forward, or rewind the tutoring tape.

Finally, the Solo Solvers in the Studying Alone condition were given instructions similar to the Collaborating condition with the exception that they would work alone. Like the Lone Observers, the Solo Solvers were also instructed that they could talk aloud whenever it felt natural to do so.

All participants were informed that the purpose of this (intervention) phase was to learn how to solve kinematics problems and that their understanding and ability to individually solve similar kinematics problems would be assessed later with new problems. They were also told to limit their problem-solving time to two hours.

After the intervention, each participant took the post-test individually, without access to the text, with an average delay of 7.4 days (ranging from 6.0 – 8.3 days). There were no significant differences between groups on this measure of delay. Thus, the post-test measured long-term learning and retention. Each student was given unlimited time to solve all four of the post-test problems.

3.2   Learning Results

The analyses in this section compare the learning performance for all five conditions, as well as explore other contrasts to either test the *active/constructive /interactive observing hypothesis*, or to set the stage for additional analyses to test the tutoring hypotheses in the second half of this paper.

3.2.1   Learning from the Pre-test to the Post-test

Pre-test and post-test problem-solving solutions were scored in terms of the steps as described in the Method section. Recall that a step could be *shallow* or *deep*, and *matched* or *unique*. Learning from the pre-test to the post-test can be analyzed as an analysis of covariance (ANCOVA), using all the matched and unique post-test steps as the dependent variable and all the matched and unique pre-test steps as the covariate. Alternatively, learning from pre-test to post-test can be analyzed as an analysis of variance (ANOVA), using the gain scores on the matched steps only. We will report both analyses, beginning with the ANCOVA using all the matched and unique steps for shallow and deep knowledge separately, then provide reasons for why the remaining analyses throughout the rest of the paper focus only on the matched *deep* steps.

3.2.1.1 Analyses of covariance for all shallow steps. Recall that all the participants in all the conditions had access to the textbook during the pre-test but not the post-test. This means that performance on the shallow steps may be enhanced at the pre-test, since they could copy the equations and notations correctly from the text. Consequently, it is important that we analyze shallow and deep steps separately. Accordingly, an ANCOVA was carried out, using all 9 of the matched and unique post-test *shallow* steps, while controlling for all 8 matched and unique pre-test *shallow* steps.

There were no significant differences across conditions in the proportion of shallow steps learned. The lack of differences among the conditions supports our interpretation that shallow steps, assessing the use of conventions, accuracy of calculations, and appropriate use of units, can be copied from the text during the pre-test so that no substantial learning differences were detected from the pre-test to the post-test. The average shallow post-test score, adjusted for pre-test score, ranged from a low of 49% for the Tutees to a high of 61% for the Solo Solvers. The lack of detectable differences in learning shallow knowledge suggests that the remaining analyses should focus only on deep steps.

   3.2.1.2 Analyses of covariance for all deep steps. Using all 22 post-test matched and unique *deep* steps adjusted for all 19 pre-test matched and unique *deep* steps, an ANCOVA shows that there were significant differences across conditions, $F(4, 64) = 2.596, p = .044$. Figure 2a shows the adjusted post-test means with standard error bars. However, there were no significant differences among the top three conditions: Tutoring, Observing Collaboratively, and Collaborating.

   Although the top three conditions were statistically equivalent, their effect sizes, contrasting each with the most conventional Studying Alone condition as a control, show a systematic decrement, from Tutoring (Cohen's $d = .815$), to Observing Collaboratively ($d = .613$), to Collaborating ($d = .326$). Furthermore, these top three conditions were clearly better learning methods since the average of their adjusted post-test scores was significantly better than either Observing Alone ($F(1, 66) = 5.111, p = .026, d = .532$) or Studying Alone ($F(1, 66) = 6.448, p = .044, d = .522$).

Insert Fig. 2 about here

3.2.1.3 Analysis of variance of matched deep steps. The preceding analyses using ANCOVA compromise our power because we have a small number of participants (we lose a degree of freedom). Hence, due to the availability of the text at the pre-test thus contaminating the learning of shallow steps, and due to the loss of one degree of freedom when using ANCOVA, henceforth, analyses of variance, comparing pre-test to post-test using *matched deep steps* only*,* will be reported throughout this paper.

Figure 2b shows the proportion and standard error bars of matched deep steps scored correctly on the pre-test and post-test. There are five patterns of results to note. First, as in the ANCOVA for all deep steps (Fig. 2a), there was a significant overall difference in the gains across conditions for matched deep steps, $F(4, 65) = 3.787$, $p = .008$.

Second, the pre- to post-test gains were significant for only three of the five conditions: Tutoring (21.3%, $F(1, 9) = 29.277$, $p = .0005$), Observing Collaboratively (17.03%, $F(1, 19) = 17.385$, $p = .001$) and Collaborating (7.19%, $F(1, 19) = 7.432$, $p = .01$). The gains for Observing Alone (8.91%) and Studying Alone (3.39%) were not significant. The effectiveness of the top three conditions is consistent with the ANCOVA results reported above, showing that the top three conditions were equivalent and better than the other two conditions. One interpretation of this result is that the three significant conditions are the only ones involving active interaction by participation in dialogues, supporting the generality of the *active/constructive/interactive hypothesis,* in that it is not restricted to the benefit of interactivity in an observing context.

Third, although the top three conditions were more similar to each other and provided significant learning, their effect sizes (expressed in terms of Cohen's *d* for the gains) descended from 1.063 for Tutoring, to 0.883 for Observing Collaboratively, to 0.405 for Collaborating, in the same ordering as the effect sizes of these three conditions in the ANCOVA of all deep steps, as well as the results in the literature. The effect sizes for Observing Alone and Studying Alone are 0.291 and 0.296, respectively.

Fourth, although the pre- to post-test gains were significant for all three conditions involving interactive dialogues as mentioned above, and descended from the Tutoring to the Collaborative Observing to the Collaborating condition, there was no significant difference between the top two conditions using ANOVA. The statistically equivalent gains of the two highest-performing conditions—Tutoring and Observing Collaboratively, is further validated in that the gains of these two conditions were significantly greater than the gains of all the other three conditions: Collaborating, Observing Alone and Studying Alone, $F(1, 67) = 6.927$, $p = .002$; $d = .556$. (Although the ANCOVA results show the top three conditions to be equivalent, whereas the ANOVA results show only the top two conditions to be equivalent, it may be that the ANOVA results provide a more sensitive comparison given our small sample size.) This equivalence of the top two conditions shows that Observing Collaboratively is as effective a learning method as the gold standard of being tutored, supporting the *active/constructive/interactive observing hypothesis,* in that it shows that observing can be as effective a way to learn as tutoring if the participants get opportunities to be active/constructive. One can be active/constructive by interacting with a peer, and not necessarily with a tutor.

Finally, Observers who observed collaboratively had marginally greater learning gains than those who observed alone, $F(1, 65) = 3.842, p = .055; d = .425$. This contrast can be accounted for by the interactions between the paired participants afforded by the Observing Collaboratively condition as compared to the Observing Alone condition, thus directly supporting *the active/constructive/interactive observing hypothesis.*

3.2.2 Good versus Poor Tutees

The overall pretest scores of the Tutees were highly variable, suggesting that some Tutees independently learned more about the material covered in Chapter 5 than other Tutees. Their pre-test scores tended to cluster into two equal groups, with the top half showing much higher scores than the bottom half, $F(1, 8) = 13.94, p = .006. d = 2.360$. Therefore, the 10 Tutees could be divided into 5 Good and 5 Poor Tutees on the basis of a median split on their pre-test scores. Because the pre-test assessed what the Tutees have learned on their own from studying Chapter 5, the pre-test is not a measure of prior background knowledge because all Tutees had a similar lack of background knowledge in physics. Instead, the pre-test is a measure of the Tutees' ability to learn without the intervention of the Tutor. Thus, the Tutees' pre-test scores can be conceived of as an index of whether they were good or poor learners.

The Good and Poor Tutees' learning ability could be further differentiated in the time and errors in learning the materials prior to tutoring. Although not statistically significant, it did take the Poor Tutees more time ($M = 440.60$ minutes) to study Chapters 1-5 than the Good Tutees ($M = 388.00$ minutes). Furthermore, even though all Tutees were trained on the first four chapters to a criterion of 80% on problem solving, the Good Tutees did make significantly fewer errors ($M = 3.0$) than the Poor Tutees ($M = 6.2, F(1,$

8) = 5.885, p = .041, d = .609). In short, the Good Tutees were better learners, in terms of how much they could understand in a shorter amount of time.

Differences between the two groups can also be detected during tutoring. While being tutored, the Poor Tutees generated many more errors across all three tutoring problems on average (M = 89.2) than Good Tutees (M = 55.8, $F(1, 8)$ = 27.474, p = .001, d = 3.276). Moreover, Poor Tutees expressed confusion twice as frequently (M = 6.8) as the Good Tutees (M = 3.1, $F(1, 8)$ = 17.551, p = .003, d = 2.644). Being good learners translated to more learning as measured by the pre- to post-test gains. The Good Tutees had greater learning gains (24.6%) than the Poor Tutees (15.6%) on matched deep scores, $F(1, 8)$ = 9.502, p = .015, d = 1.976.

Overall, these measures confirm the fact that a median split of Tutees on the basis of their pre-test performance is a legitimate way to contrast more successful learners (the Good Tutees) with the less successful learners (the Poor Tutees). This contrast serves a useful purpose for subsequent analyses.

3.2.3   *Collaborative Observers Learn More from Observing Good versus Poor Tutees*

The 10 Collaborative Observers who observed the 5 Good Tutees' tapes (henceforth referred to as the Good Tapes) learned substantially more (M = 21.9%) than the 10 Collaborative Observers who observed the 5 Poor Tapes (M = 7.8%; $F(1, 18)$ = 6.723, p = .018, d = 0.573). In fact, the 21.9% gain of the Collaborative Observers of Good Tapes is comparable to (in fact slightly exceeds) the mean gain for the Tutees per se (21.3%).

The Collaborative Observers themselves can also be divided into Good and Poor Observers, based on a median split of their pre-test scores. Although the Good Tapes

were more effective than the Poor Tapes in general for both the Good and Poor

Observers, the advantage of the Good Tape is much more pronounced for the Poor

Observers (see Figure 3). In fact, the Poor Observers ($n = 6$) benefited the most from

observing the Good Tapes, in that they gained significantly more than the other three

groups of collaborative Observers combined ($n = 14$, $F(1, 18) = 6.907$, $p = .017$, $d =$

1.153). They gained on average 31.3%, which far exceeded the average gain of even the

Good Tutees (24.6%). It does not appear as though the Good Observers were performing

at ceiling given that they only scored 76% correct on the post-test. Hence, the Good

Tapes were extremely effective for learning from observing, especially for Poor

Observers. Despite the small sample sizes, this finding is extremely encouraging as it

indicates the tremendous potential of providing good tapes for poor learners to observe.

Insert Fig. 3 about here

3.2.4   *More Interactions Facilitate Learning from Observing*

Although this paper is not about how the Observers learn *per se* (such analyses

are being carried out and will be reported in a later paper), we can further test the

active/constructive/interactive observing hypothesis by showing that how interactive the

Observers were is an important determinant of their learning. One simple way to measure

how interactive the Collaborative Observers were with each other while watching the

tapes is to compute the amount of substantive contributions (to be defined in the latter

half of this paper) made by each partner in their exchanges.

The Collaborative Observers did engage in a substantial amount of interaction producing in the range of 274 to 492 exchanges (except for an outlier pair that had only 22 exchanges). We computed the amount of interactions between each pair by the difference in the proportion of substantive contributions each partner made. A difference of less than 15% (with an average difference of 9.2%) was considered High Interactive Observers since both partners were contributing more equally to problem solving, and there were 6 pairs of such dyads. The remaining four pairs had a difference of greater than 15% and can be considered disproportionately Low Interactive Observers. The High Interactive Observers learned significantly more ($M = 26\%$) than the Low Interactive Observers ($M = 4\%$, $F(1, 18) = 10.313$, $p = .005$, $d = 1.506$). Thus, this result further validates our *active/constructive/interactive observing hypothesis*.

3.2.5   Active Lone Observers

If being active/constructive/interactive is the key to learning, then another way to test this hypothesis is to see if we can tease apart those Lone Observers who were active versus those who were more passive. The *active/constructive/interactive observing hypothesis* predicts that the more active individuals would learn more. Accordingly, for the Observing Alone condition, four participants were identified as being more active in the following ways. Contrasting these four participants with the other six, these active Lone Observers were more likely to write problem-solving steps on the whiteboard ($M = 74.00$ steps vs. $M = 7.33$); to manipulate the tutoring tape by pausing, rewinding, and fast forwarding ($M = 29.75$ vs. $M = 19.33$); and to pose self-querying questions such as, "Mass A was cancelled. Are you sure you're allowed to do that?" ($M = 7.5$ vs. $M = 0.2$).

Not surprisingly, these four active Lone Observers gained more (18.4%, from pre-test to post-test on their matched deep step scores) than the six passive Lone Observers (2.6%), even though the two groups did not differ in terms of their pre-test, deep step scores (34.7% and 35.8%, respectively). Note that the 18.4% gain of the active Lone Observers is comparable to the average gains of the Tutees (21.3%) and the Collaborative Observers (17.0%). Moreover, the 2.6% average gain for the more passive Lone Observers was comparable to the gains of the Solo Solvers (3.4%). Even though the number of active Lone Observers is small, the differential gains comparing the active Lone Observers versus the passive Lone Observers lends further support to the *active/constructive/interactive observing hypothesis.*

3.3   Summary

The results reported in this section contrast the advantages of learning from tutoring, from vicariously observing tutoring (either collaboratively or alone), from collaborating, and from studying alone as our baseline control condition, since it mimics most closely what students can acquire without any help from either a tutor, a peer, or a videotape. There are several findings to note. First, our data from the Tutoring, Collaboratively Observing and Collaborating conditions replicate the evidence in the literature. As discussed in the Introduction, learning gains from tutoring typically have a larger effect size (.4 – 2.0) than those from observing collaboratively (1.12) and collaborating (.21-.88). The learning gain effect sizes we found, of 1.063 for Tutoring, 0.883 for Observing Collaboratively, and 0.405 for Collaborating, follow the same trend, and are significantly better than that of Studying Alone (effect size of 0.296). Note that

the magnitude of our effect sizes reflects the fact that we are measuring learning of deep knowledge in a fairly difficult problem-solving domain.

Second, the literature is often discrepant on the learning effectiveness of observing vicariously, in the sense of learning from observing someone else learns. Our hypothesis, that the discrepancy in the literature may arise from how active/constructive/interactive an observer is, is supported in four ways. The first comparison that supports this *active/constructive/interactive observing hypothesis* is the equivalent learning gains of the Collaborative Observers and the Tutees. This null effect suggests that having and using the opportunity to interact with a peer is as effective as directly interacting with a tutor. The same interaction interpretation can also be offered for the second comparison, between the Collaborative Observers versus the Lone Observers. The better learning outcomes of the Collaborative Observers may have arisen from their having the opportunity to be more active and constructive, by interacting with their partners. The third result that supports our hypothesis is the contrast between the more interactive Collaborative Observers and the less interactive Collaborative Observers. The more interactive ones learned more. The fourth comparison that supports our hypothesis can be gleaned from teasing apart those Lone Observers who were more active versus those who were more passive. The more active Lone Observers learned more.

In sum, these four findings support our *active/constructive/interactive observing hypothesis,* suggesting that *collaboratively observing tutoring while solving problems*, may be an effective learning environment. Moreover, the greater gain of the active Lone Observers poses the plausible interpretation that the advantage of the three conditions in

which students could participate in dialogues may not have to do with dialogues *per se*, but rather, that participating in dialogues is a natural forum for being more interactive and thus constructive.

Although we consider the Studying Alone condition to be one of our poorest baseline control condition, we should point out that this Studying Alone condition is analogous to what the literature has referred to as learning by doing (Anzai & Simon, 1979). Although learning by doing or problem-based learning has been promoted in the past as an optimal way to help students learn, especially in medical education, more recent evidence shows that learning by doing is not as effective as learning from studying worked-out examples (Sweller, van Merrienboer & Pass, 1998), consistent with our comparisons show that learning by doing is definitely not as effective as the top three conditions. Okita and Schwartz (2006) also have some intriguing evidence consistent with ours, showing that learning by doing (such as answering questions) is not as effective as learning by observing someone else answers questions.

Although our Observing Collaboratively learning environment is not novel, we have provided systematic data comparing it with other learning methods. For example, such a learning environment has been explored in a single case study by Frederiksen, Donin, Meilleur, Roy, and Bracewell (1999), but they did not contrast this learning environment with other learning environments to assess its benefit. Such a learning environment was also explored by Craig et al. (2004). They compared learning from interacting with an animated tutor agent, with learning from either observing tutoring individually (Exp. 1) or collaboratively (Exp. 2). Although their tutoring condition was superior to their individual observing condition (consistent with our Tutoring vs.

Observing Alone contrast), their collaboratively observing condition was not significantly different from either their tutoring condition or their observing alone condition.

The discrepancy between the null results reported by Craig et al. (2004) in Exp. 2 (of no overall significant differences across their three conditions) and ours may be due to the fact that their collaborative observers interacted infrequently, on average 2.91 times per 35-minute tutoring session, whereas our Collaborative Observers interacted on average 15.36 episodes per 35-minutes duration. Thus, being not as interactive may have weakened the learning gains of their collaborative observers, as the contrast in learning gains of our High Interactive and Low Interactive Collaborative Observers shows. Nevertheless, their results, even though not significant, could be viewed as compatible with ours. The effect sizes for their tutoring (2.06) was similar to the effect size for their collaboratively observing condition (2.17), which in turn was much better than their observing alone condition (0.97). In short, the ordering of their effect sizes for those three conditions in fact matches the pattern of our effect sizes completely.

4.   Three Tutoring Hypotheses and Analyses of the Tutoring Protocols

In this second half of this paper, the three tutoring hypotheses proposed in Chi et al. (2001), differentiated globally in terms of whether the tutor, the tutee, or their interaction, is responsible for the benefits of learning from tutoring, will be further elaborated and evaluated. In order to differentiate among the three tutoring hypotheses, an added twist in the interpretation of our data is to look at how the tutoring dialogues affect the learning of the Collaborative Observers. Thus, this section of the paper reports analyses of the tutor-tutee dialogues from the benchmark Tutoring condition in order to

further understand how the Tutees and the Collaborative Observers learn. This entire section therefore concerns the Tutees and the Collaborative Observers only.

4.1 Three Hypothesis about Tutoring Effectiveness

Three general hypotheses have been proposed for the learning benefit from being tutored (Chi et al., 2001). The first hypothesis, referred to as *the tutor-centered pedagogical hypothesis (*or in brief, *the tutor-centered hypothesis)*, assumes that learning from tutoring is enhanced because the tutor undertakes pedagogical *moves* (such as explaining, scaffolding, giving feedback, or motivating) that are tailored to the tutees. This hypothesis similarly underlies research that examines effective and ineffective teaching practices (Shulman, 1986), in that it assumes the effectiveness of teaching affects student learning. Such an assumption may implicitly underlie tutoring research that searched for tutors' optimum pedagogical strategies and moves (Hume, Michael, Rovick, & Evens, 1993; Merrill, Reiser, Ranney, & Trafton, 1992; Merrill, Reiser, Merrill, & Landes, 1995; Evens, Spitkovsky, Boyle, Michael & Rovick, 1993; Putnam, 1987; Sleeman, Kelly, Martinak, Ward, & Morre, 1989; Lepper, Wollverton, Mumme, & Gurtner, 1991; VanLehn et al., 2003).

Crediting a tutor almost entirely for a tutee's learning is rational, but actually without basis in evidence. Perhaps the hypothesis arose from three observations: (a) that powerful techniques are used by a few unique and exceptional orators and tutors such as Socrates (Collins & Stevens, 1982; Collins, Brown & Newman, 1989); (b) that a tutor is generally knowledgeable about the content domain that she is tutoring, therefore this domain expertise is confounded with the assumption that the tutor must also be an expert on the pedagogy of tutoring; and (c) that a tutor does typically control, lead, and

dominate the tutorial conversation (Graesser, Person & Magliano, 1995; Chi et al., 2001).

Granted one can determine that some tutorial moves might affect learning (specifically

feedback can accelerate learning in the context of problem solving, Anderson, Corbett,

Koedinger & Pelletier, 1995), these three observations may be epiphenomena, in that the

differential tutoring moves may not in fact be responsible for tutee's learning.

We can evaluate this *tutor-centered pedagogical hypothesis* in terms of three

components: *frequency, quality,* and *adaptiveness*. That is, for the first component, if we

assume that a tutor's move is responsible for a tutee's learning, then the more often such

a pedagogical move is undertaken, the better the tutees ought to learn. Our prior results

found no support for this *frequency* component in terms of correlations. For instance, we

found that although our novice tutors explained frequently to tutees, the tutees did not

seem to learn much from these explanations (Chi et al., 2001, Study 1).

The *quality* component was not examined directly in our prior study. One could

argue that expert tutors might give excellent explanations that novice tutors could not,

assuming that good instructional explanations facilitate learning much more so than poor

instructional explanations (Eisenhart, Borko, Underhill, Brown, Jones, & Agard, 1993).

Since the tutoring dialogues in this study involve an experienced teacher/tutor, we can

infer the *quality* component by analyzing the frequency component again. If the same

results are obtained here with this more expert-like Tutor as our prior results with novice

tutors, then the evidence indirectly tests the quality component.

The third component of *the tutor-centered hypothesis* is *adaptiveness*.

Adaptiveness is a very complex concept. It can refer to a tutor's selection of the

appropriate moves, delivered at the right moment, and based on a tutee's need for

feedback and help (Murray & VanLehn, 2005). Thus, being adaptive can be operationalized to mean that a tutor (a) must *choose* the appropriate moves (or problems to be solved by the tutee) that are tailored to the tutee, knows (b) *when* to deliver his feedback, explanations, and scaffolding hints (such as contingent upon the correctness of a tutee's response, Wood, Wood & Middleton, 1978), and this knowledge must be gleaned from his continuous (c) *assessment* of the tutee's competence and understanding. Moreover, *assessment* can mean from either a normative perspective or from the tutee's (or student's) perspective. For example, we found that inexperienced tutors could not assess tutees' deep understanding accurately, from the students' perspective (Chi, Siler, & Jeong, 2004); whereas tutors are usually quite capable of assessing tutees' competence from the normative perspective (Putnam, 1987). In short, there is scant evidence to test the many aspects of the adaptiveness component of the *tutor-centered hypothesis*.

A second hypothesis for the benefit of tutoring is the idea that a tutoring context, by definition, is one in which a tutee has greater opportunities to have a one-on-one dialogue with the tutor, as compared to a standard classroom context. This opportunity to be constructive potentially could cause the tutees to learn more. We called this *the student-centered constructive hypothesis* (or *the student-centered hypothesis),* to contrast the role of the tutees from the role of the tutor. We provided preliminary evidence in support of this hypothesis. For example, when the tutors were suppressed from giving any explanations and feedback at all, and could only give prompts (Study 2, Chi et al., 2001), the tutees learned just as effectively as when the tutors gave a substantial amount of explanations and feedback (Study 1, Chi et al., 2001). We attributed the tutees' learning to the constructive responses that they gave to tutors' scaffoldings.

Finally, a third proposed hypothesis is *the interactive coordination hypothesis* (or *the interaction hypothesis),* which states that tutoring effectiveness depends upon the joint or coordinated effort of both the tutor and the tutee. For example, our evidence showed that some tutor moves (such as scaffolding) were more beneficial for tutees' learning than other tutor moves (such as giving explanations; Chi et al., 2001, pp. 499-500). Moreover, when we encouraged the tutors to do more scaffolding than explaining, this resulted in an increase in the number of multi-turn *deep* tutor-tutee interactions (see Fig. 9, Chi et al., 2001). Both of these results suggest that scaffolding can elicit more meaningful and elaborate joint construction. Although we could infer that these two results confirmed that some kind of interactions between a tutor and a tutee contributed more toward learning than other kinds, it was actually impossible to isolate the contributions of the tutees independently of the contributions of the tutors. That is, we could not discriminate whether the learning arose from the tutors' scaffoldings *per se* or from the tutees' constructive responses *per se* or from their interactions.

In sum, our previous studies (Chi et al., 2001; Chi et al., 2004) provided sufficient evidence to question the *tutor-centered hypothesis* and to highlight both the *student-centered* and the *interaction hypotheses* as potential accounts for tutees' learning. The current study hopes to provide additional evidence to support and/or refute these hypotheses, using a procedural domain (problem solving) rather than a conceptual domain (human circulatory system), a single more expert-like tutor rather than multiple novice tutors, and college students rather than middle-school students as tutees. Moreover, by inferring the effectiveness of tutoring from the additional perspective of how the Collaborative Observers learn, we might gain better insight into all three

hypotheses since the tutoring was not tailored to the Observers, nor were the Observers constructing responses to the Tutor, and moreover, the Observers heard not only what the Tutor said, but they also heard what the Tutees said.

4.2 Segmentation and Grain Size

Learning studies using complex materials, such as solving physics problems, can generate a massive amount of protocol data. Because of the labor-intensiveness of transcribing and coding such a massive amount of data, we tested a short-cut method and undertook duplicate coding for 20% of the data for the purpose of calculating inter-rater reliability. However, we did inject several "validating" analyses to gain further confidence in our coding. By this, we mean alternative codings, often at a different grain size, or with a different set of goals, to see whether the results from different codings replicate each other, or whether some codings replicate robust evidence in the literature.

We begin by explaining how the 10 tutoring videotapes were segmented in a short-cut way after they were transcribed. Each transcript was first segmented according to speaker *turns*. A turn, by definition, is speech by a single speaker (Traum & Heeman, 1997). Then the transcript was further segmented according to either the speaker's *intonation* (e.g., a falling tone, a rising tone), *pauses,* or *changes in action* (e.g., from talking to writing on the board, or from reading the problem statement to writing on the board). Two coders independently segmented 20% (2 of the 10) tutoring video transcripts while watching the tutoring videos, and agreed on 2466 (97.03% of the total) segments.

This short-cut method of segmentation, based on the structure of speech (*turns, intonation, pauses,* and *changes in action*) can be carried out much more objectively and rapidly than segmentation based on an analysis of the content of speech, as we have

routinely done in the past (Chi, 1997; Chi et al., 2001). Therefore, it is important to know whether segmentation based on the structure of speech is adequate in comparison to segmentation based on the content, since the former can be more easily automated. To verify this, we selected the middle 20% of each of the ten tutoring transcripts and coded according to the content method of segmentation (Chi, 1997). The segment boundaries were then compared across the two methods, yielding concordance rate of 89.1% for the 2,416 coding decisions made across the two systems. Not only does this result indicate a high level of agreement between the two methods, but the percentage of disagreements between the two methods was quite symmetric (5.4% of the time a segment was indicated by the content coding but not according to the structure coding; and 6.1% for the alternate instance), indicating that there was no systematic bias one way or the other in terms of segmenting at a consistent grain size.

Using the structure rather than the content of speech, a "segment" does nevertheless roughly correspond to what we have previously referred to as a "statement"; that is, to a single idea, presented by a single speaker within a turn. Thus, a *segment* is our smallest unit of analyses. The utterances made by the Tutor, shown below, were coded as three segments, the boundaries are shown by the double lines.

"So you have therefore written the equation of motion.//

And from using the equation of motion you have been able to find out what would be the normal reaction for block A.//

Now similarly there is a equation of motion for block B.//"

Segments can also be combined into interactive *dialogue units* when tutor-tutee responses are considered jointly, often in adjacent pairs of turns. Because a turn can

contain multiple segments, the last segment within a turn (from the Tutor, for example) can be analyzed with the response in the next turn (by the Tutee), to create a dialogue unit. Finally, segments can also be combined into an *episode* unit. An episode corresponded to the consecutive talk and problem-solving segments that referred to the same problem-solving node, from the model of problem solutions (see Fig. 1 for an example of a problem solution model).

The results will be reported below in three sections, corresponding to the three grain sizes: *segments* within one turn, *dialogues* that are consecutive segments involving two turns, and *episodes* often involving multiple turns. The details of the codings for each grain size will be unpacked as each result is being described.

4.2.1   Independent Segment Analyses

The Tutor, on average, made a total of 686 segments whereas the Tutees averaged 443 segments, per tutoring session. This approximate 3:2 tutor-tutee ratio is typical in tutoring, in which the tutor usually dominates the conversation by talking more (Chi et al., 2001; Graesser & Person, 1994). In our prior study, the tutor-tutee ratio was even more pronounced (621 tutor statements vs. 206 tutee statements, roughly a 3:1 ratio). The smaller difference in this current study may be attributed to the experience of this Tutor in the context of our research, in that he recognized the advantage of making fewer statements himself and eliciting more responses from the Tutees. One could perhaps take the ratio of tutor-tutee contributions as an index of tutor expertise, suggesting that our Tutor in fact is more of an expert tutor.

Overall, the tutorial dialogue, considering both the Tutor and the Tutees' moves, was far more extensive with the Poor Tutees (1393 mean number of tutor-tutee segments)

than with the Good Tutees (864 segments). An obvious interpretation of this result is that the Poor Tutees needed more help. This overall greater amount of dialogue with the Poor Tutees translated into a greater frequency of all types of tutor moves; therefore, it is sometimes more appropriate to calculate the proportion and other times to use the frequency of moves in the analyses to be reported below.

4.2.1.1 Learning from the Tutor's moves? Tutor moves were defined as instructional segments, uttered by the Tutor, that were relevant to the pedagogical task of tutoring. Tutor segments were categorized as either an explaining move, a scaffolding move, a feedback move, or various other miscellaneous moves (such as summarizing, comprehension checking, tutor responding to tutee questions, false starts, and so forth). Tutor moves for the middle 20% of each of the ten tutoring protocols were independently coded by two raters into these four categories. Based on the Kappa coefficient, the inter-rater reliability for this coding indicated substantial agreement ($\kappa = 0.758$). Subsequent analyses will focus on the three largest categories: explanations, scaffolding, and feedback segments only.

An explanation is an utterance in which the Tutor defines physics concepts or principles, provides an interpretation of important problem situation features, describes how to carry out a particular procedure, the conventions used to carry out a particular mathematical or physics problem-solving step, or the outcome of applying some procedure. Below is an example of a tutor explanation, generated in one turn, containing four segments:

If there is a net force F, then there will be an acceleration, a, on that object//

If there is no force, then there is no acceleration.//

So this is the equation which tells you that whether an object -an object is

accelerated or not. //

And umm, therefore, it is called the equation of motion.//

A scaffolding is defined as either a prompt that is content-free (superficially, it

gives away no information) or some kind of support for helping or guiding the tutees

toward understanding. The support can take the form of a hint, an assertion with an

expectation to fill-in-a-blank, a direct or indirect question, and so forth. In Chi et al.

(2001), we identified 14 different forms of scaffolding. Below are three examples of

scaffolding taken from different contexts in the protocols, each of one segment length:

Weight is the? … //

Acceleration due to?… //

When a force acts on body, uh, how does the body react to it?//

A feedback segment can be either a short *positive* (e.g. "right") or *negative* (e.g.,

"no, no") response about the correctness or incorrectness of what the Tutees said or did,

or it can be more extensive, in terms of *correcting* what the Tutees did incorrectly (e.g.,

"No, the Earth") or *elaborating* further on what the Tutees did or stated (e.g., "No, it

should be accelerating towards A and B"). In the latter two cases, the Tutor's feedback

would be coded only as a corrective/elaborative feedback (and not double-coded as both

a negative and a corrective/elaborative feedback). These feedback segments can be given

to either correct or incorrect Tutee responses.

The top section of Table 3 shows the average number of segments per session and

their proportion for each type of Tutor's instructional move. Note that scaffolding is the

largest category of the Tutor's instructional moves, consisting of 36% of his total

statements. In contrast, in our prior study, scaffolding consisted of only 5% of the total

number of instructional moves. Likewise, explanations consisted of 23% of the total

instructional moves, whereas in our prior study, the tutors' explanations consisted of 53%

of the total statements (Chi et al., 2001, Fig. 2). The reversal in the ratio of explanations-

to-scaffolding might be caused by the expertise of the current Tutor, whereas our prior

study involved 11 novice tutors. In fact, one might consider the ratio of explanations to

scaffolding moves as another index of our Tutor's pedagogical expertise.

Insert Table 3 about here

Did the frequency of Tutor moves correlate with Tutees' learning? No. There

were no significant correlations between the average number of Tutor's explanation nor

scaffolding segments *per se* with either the Tutees' or the Collaborative Observers'

matched deep step gains (see the last 2 columns of Table 3, top). This result replicates

what was found in Chi et al. (2001). There, neither the tutors' explanations (Table 3) nor

scaffoldings (Table 3, Model 2), correlated with deep learning. That is, if we extract from

the protocols only the Tutor's moves in terms of the frequency of explanations and

scaffoldings that the Tutor provided, then receiving and hearing those moves as

independent monologues did not have an impact on either the Tutees' or the Observers'

learning.

What about the Tutor's feedback? As shown in Table 3, there was a negative

correlation between all types (positive, negative, corrective, elaborative) of Tutor's

feedback and the Tutees' learning ($r = -.603$), but no correlation with the Observers'

learning, and the negative correlation with the Tutees' learning was marginally

significant ($p = .065$). This marginal, negative correlation between the Tutor's feedback

and the Tutees' learning cannot be mediated by Tutees' errors because there is not a

significant correlation between Tutees' errors and Tutees learning. That is, it is not the

case that the more errors a tutee makes (thereby eliciting more tutor feedback), the less

they are likely to learn, thus accounting for the negative correlation. This puzzling

negative correlation will be examined more closely later.

4.2.1.2 Learning from the Tutees' moves? The preceding section analyzed the

Tutor's moves independently of the Tutees' responses. Such an analysis, in essence,

treated the Tutor's moves as instructional monologues, and there were no significant

correlations between the frequency of the Tutor's moves with the Tutees' nor the

Observers' learning (except for a marginal negative correlation with feedback for the

Tutees only). If neither the Tutees nor the Observers learned by considering the Tutor's

independent moves, then how did they learn? In this section, we analyze the Tutees'

independent learning moves.

Tutees' segments were categorized as either a substantive or a non-substantive

learning move in terms of the content, regardless of the form of the segment, such as an

assertion or a question. A substantive segment is defined as a meaningful contribution to

an ongoing activity, such as problem solving, or a relevant response to the Tutor's

explanations. For example, to the Tutor explanation shown below the Tutee's response

would be coded as a substantive one:

TUTOR: See this equation is true for constant acceleration.//

Now the acceleration is constant here.//

Forces are not changing on the weight so the acceleration is constant.//

TUTEE: The initial velocity is zero then.//

A non-substantive segment is defined as a continuer, a repetition, an agreement, or off-task remarks. To the Tutor's explanation shown above, if the Tutee had responded with "alright", then that would be coded as a non-substantive response.

The middle section of Table 3 showed that 52% of the Tutees' segments were substantive. The correlation of substantive moves with matched deep step gain was $r = .605$ and it approached significance ($p = .064$); whereas the correlation of non-substantive moves with matched deep step gain was strongly negative ($r = -.899, p = .000$). Thus, the Tutees learned only when they responded with substantive contributions, but they definitely did not learn when they constructed non-substantive responses, again, replicating our previous results (Chi et al., 2001). Thus, being responsive *per se* is not sufficient; one must construct *substantive* responses in order to learn.

Substantive segments can be further divided into those that are relevant or irrelevant. Relevant substantive segments are those that are responsive to the Tutor's comments in the sense of building-on or following-up to the Tutor's comments. The following underlined segment would be an example of a relevant substantive response:

TUTOR: If I push it, it's, velocity becomes some-something.//

TUTEE: Mm hmm. [tutee nods yes]//

TUTOR: So from zero to something, there is a change.//

TUTEE: Ok, so yeah.// It wouldn't be a constant.//

Irrelevant responses are those that are not responsive to the Tutor's comments; but are nonetheless substantive. The underlined example below is an example of an irrelevant but substantive response:

TUTOR: It seems reasonable?//

TUTEE: That the Earth is accelerating.//

TUTOR: Because of these masses.//

TUTEE: [tutee laughs] No. Those are some pretty big masses.//

Tutees benefited from constructing substantive responses, both relevant (the correlation is $r = .641, p = .047$) as well as irrelevant ones ($r = .620, p = .056$). (See Table 3, bottom.) The Observers, however, seemed to be able to benefit somewhat only from overhearing the relevant responses (trend, $r = .398, p = .082$).

The fact that the Tutees could benefit from making substantive responses whether or not they were relevant, replicates our overall self-explanation effect (Chi et al., 1994; McNamara, 2004) if we assume that irrelevant substantive responses are analogous to idiosyncratic self-explanations. In particular, in Chi et al. (1994), we claimed that students could learn whether they generated correct or incorrect self-explanations. A simple interpretation for this latter finding is that, although the substantive responses may seem irrelevant from the normative perspective, they can be conceived of as self-explanations that serve the Tutees' own purposes of repairing and refining their own understanding (Chi, 2000). This same interpretation can be used to explain the modest benefit Observers had from overhearing the relevant substantive follow-ups but not from overhearing the irrelevant responses, because the irrelevant ones would not make sense to

an observer, since they served the Tutees' own purpose of repairing and refining their own understanding.

This result, that Tutees learned from constructing substantive responses, further reinforces our previous interpretation of Study 2 in Chi et al. (2001). There, the tutees seemed to have learned in an artificial tutoring condition in which the tutors were suppressed from explaining but encouraged to scaffold. The tutees in the suppressed condition learned just as well without tutors' explanations but with many more tutor scaffoldings (see Fig. 8, Chi et al., 2001). We had inferred then that the tutees must have learned from the benefit of constructing responses to the tutors' scaffoldings, and the result provided here further confirms that interpretation, along with some evidence in VanLehn et al. (2003, Tables 11 and 12), Litman and Forbes-Riley (2006), and Jackson, Person and Graesser (2004).

4.2.1.3 Summary of segment-level analysis. Overall, the pattern of correlation results shows that the frequency of the Tutor's moves had mostly no effect on the Tutees' (nor the Observers') learning, replicating our tentative results from the 2001 study. Moreover, the Tutor's feedback moves were somewhat detrimental to the Tutees' learning. Thus, these results do not support the *frequency* component of the *tutor-centered hypothesis*. Moreover, because our Tutor is an experienced teacher, the lack of correlation between his instructional moves and Tutees' learning suggests indirectly, that the *quality* component of the *tutor-centered hypothesis* is not supported either. Thus, there was no support for two of the three components of the *tutor-centered hypothesis*.

The Tutees' moves, however, did affect their own learning, but only if they constructed substantive responses (whether relevant or irrelevant). This finding, that the

students' own construction is responsible for learning from tutoring, further supports the *student-centered constructive hypothesis*.

The Observers could not learn by overhearing either the Tutor's or the Tutees' independent moves, even if they were substantive. But there was a trend for a correlation with their learning when the Tutees' responses were not only substantive, but relevant as well, most likely because substantive relevant responses are normative, whereas substantive irrelevant responses are specific to the Tutees' own mental models only. This trend will be examined in greater detail in the next section.

4.2.2   Interactive Dialogue Analyses

Testing the *tutor-centered* and the *student-centered hypotheses* involved analyzing tutor's instructional moves and tutees' learning moves independently. Even so, we often could not discriminate whether the Tutees' learning arose from their own construction or from receiving the Tutor's instructional moves. For example, as shown in the bottom of Table 3, the Tutees learned from constructing both relevant and irrelevant substantive responses. But we could not determine whether their learning arose from their own self-directed construction only, from receiving the Tutor's instructional moves, or from some interaction of the two. However, analyzing interactive dialogue units as well as analyzing from the perspective of the Observers' learning, might allow us to differentiate the contributions of the Tutor, the Tutees, or their interactions, toward the Tutees' learning. Accordingly, in this section, analyses of the tutoring protocols will take on a larger grain size, in terms of tutor-tutee dialogue units.

4.2.2.1 Tutees' relevant substantive follow-up responses to tutor's scaffolding and explanations. Since the Observers did not learn at all from overhearing the Tutees

irrelevant responses (Table 3, bottom), we focus only on the Tutees' relevant follow-up responses. Coding of relevant Tutee responses were actually interactive codings that examined adjacent pairs of turns, because relevance must be defined in the context of the content of the prior utterance. These prior Tutor utterances were either explanations or scaffoldings. (There were no instances of Tutees making immediate relevant substantive responses to Tutor feedbacks because Tutor feedbacks were immediately followed by either some other type of Tutor move before a Tutee had a chance to respond (e.g., feedback followed by a scaffolding or explanation), by a non-substantive Tutee move (e.g. a continuer or repetition of what the Tutor said without adding any new information), or by a Tutee non-substantive response (e.g., a request directed to the Tutor; or an assertion like "I can do this math" or "Ok I see where you are going.") Accordingly, we examined tutor-tutee dialogue units that were relevant Tutee responses that either followed a Tutor scaffolding or a Tutor explanation.

As shown at the bottom of Table 3, there were on average, a total of 230 substantive segments made by each Tutee per tutoring session. Of these, only 99 of them were relevant follow-up responses to the Tutor's scaffolding or explaining moves. Here is an example of a Tutee's relevant follow-up (underlined) given to the Tutor's scaffolding (more Tutor segments are provided for context) :

TUTOR: No M is acceleration of what?//

This force is acting on what?//

TUTEE: This force is acting on- the ground.//

And here is an example of a Tutee's relevant follow-up given to the Tutor's explanation:

TUTOR: So this will be pulling this object to it.//

So since A and the force that A and B are experiencing due to G is

attractive force- force- directed toward G- //

so G should be experiencing a force which is directed toward A and B//

so that will be upward force//

TUTEE: Oh okay- so I have this backwards.//

For the Tutees, their learning correlated significantly with constructing a relevant follow-up to the Tutor's scaffolding ($r = .656$, $p = .039$), more so than constructing a relevant follow-up to a Tutor's explanations ($r = .576$, $p = .082$), although there is a trend here too (see Table 4). However, we cannot tease apart whether the Tutees learned as a result of receiving the Tutor's scaffolding and explaining pedagogical moves, or from their own construction. Since the Observers were not constructing the responses, whether or not they learned by overhearing these tutor-tutee dialogue units would help differentiate the interpretations above. Table 4 (last column) shows that the Collaborative Observers learned significantly only when they overheard *scaffolding*-relevant follow-up dialogue units ($r = .434$, $p = .056$) but not *explanation*-relevant follow-up dialogue units. Their differentiated learning outcomes suggest that the source of the Tutees' learning might be different for the two types of instructional moves. We offer the interpretation that the Tutees learned from co-construction when the Tutor scaffolded them, and they learned from their own self-construction when the Tutor explained to them. That is, when the Tutor explained, the Tutees might have learned from constructing their own relevant substantive responses (and not from receiving the explanations since there was no correlation between Tutor's explanations and Tutee learning, see Table 3 again); whereas when the Tutor scaffolded, the Tutees might have learned from joint or co-constructing

with the Tutor, in the general sense of building on to and/or extending upon (Tao &

Gunstone, 1999) what the Tutor said.

Insert Table 4 about here

Why would Tutor's scaffolding enhance joint construction more so than

explaining? Note that a scaffolding tends to be a question, a prompt or a hint, some move

that is brief and expects a follow-up response, whereas an explanation tends to be longer

and more didactic-like assertions that do not necessarily expect a response (see the two

previous protocol examples). Therefore, by its very nature of being short and

anticipatory, it is easier to understand and build-on a scaffolding move than an explaining

move. In short, one interpretation is that the short and anticipatory nature of scaffolding

invites joint construction.

One way to test our interpretation that *scaffolding*-relevant follow-up dialogue

units are jointly constructed more so than *explanation*-relevant follow-ups, is to analyze

their coherence. Since joint construction involves building-on and extending each other's

utterances, one would expect jointly constructed dialogues to be more coherent than non-

jointly constructed dialogues. To test this coherence hypothesis, we compared the

cohesiveness of scaffolding-relevant follow-up and explanation-relevant follow-up

dialogue units by using a computer tool called Coh-Metrix, which was developed for

analyzing the cohesion, language characteristics, and readability of texts (McNamara,

Louwerse, Cai, and Graesser, 2005). Applying this text analysis tool revealed that local

cohesion based on adjacent sentences for *scaffolding*-relevant follow-up dialogue units

(M = LSA local measure of cohesion of 0.28) were on average significantly more cohesive than *explanation*-relevant follow-up dialogue units (M = LSA local measure of cohesion of 0.19; $F(1,9) = 5.685$, $p = .041$; $d = 1.078$). In addition, there was a strong trend for global cohesion across sentences to be higher for *scaffolding*-relevant follow-up dialogue units (M = LSA global measure of cohesion of 0.26) than *explanation*-relevant follow-up dialogues (M = LSA global measure of cohesion of 0.16; $F(1,9) = 4.491$, $p = .063$; $d = 1.059$).

In short, the fact that *scaffolding*-relevant response units are more coherent than *explanation*-relevant response units, supports our interpretation that *scaffolding*-relevant response units were more likely to be jointly constructed. Thus, the Tutees' learning when responding to the Tutor's scaffolding may arise from joint construction, supporting the *interaction hypothesis,* whereas the Tutees' learning when responding to the Tutor's explanation may arise from self-construction, supporting the *student-centered hypothesis*.

Why might overhearing *scaffolding*-relevant follow-up dialogues also be better for the Observers' learning than overhearing *explanation*-relevant follow-up dialogue units (see Table 4 again)? The same interpretation can be applied here as well: That is, *scaffolding*-relevant follow-up dialogues tend to be more easily understood by the Observers because they were shorter and more coherent than *explanation*-relevant follow-up dialogues. We tested whether *scaffolding*-relevant response units were in fact shorter than *explanation*-relevant response units by taking the middle 20 percent of each of the tutoring protocols and calculated the number of words produced. *Scaffolding*-relevant response dialogue units averaged 30 words whereas *explanation*-relevant response dialogue units averaged 66 words. Our interpretation that shorter dialogue units

are more comprehensible by an observer is compatible to some results of VanLehn et al. (2003); they also found that shorter learning opportunities were associated with more frequent gains.

4.2.2.2 Tutor's feedback to Tutees' errors. We reported in Table 3 (in the Feedback row) that there was an overall negative correlation between the Tutor's feedback and the Tutees' learning. In order to understand this puzzling result, it may be more meaningful to examine the effect of feedback in an interactive way, such as examining only feedback that followed an error that a tutee made, presumably because it is more difficult to predict the utility of *positive* feedback, feedback given usually to making a correct step. The frequency of positive feedback may also be a function of a tutor's style (such as giving more positive feedback for motivation purposes, Lepper et al., 1991). Moreover, Tutees may not learn as much by being told that their actions were correct. Thus, it seems more informative to analyze the effect of the Tutor's negative feedback to errors only.

Although a majority of the studies in the tutoring literature discuss feedback in terms of negative ones, the choice of giving a negative feedback is not at the discretion of the tutor, since it obviously depends on whether an error was made in the first place. However, the Tutor does have control over whether the negative feedback contains only the correct answer, or whether the negative feedback also includes elaborations and justifications. In short, feedback-to-errors can take one of three forms. Besides giving a *negative* feedback saying that the response is incorrect or "No", the Tutor has the additional option of giving a *corrective* feedback in which the Tutor basically gave the correct answer, such as:

TUTEE: [Tutee writes * g] Times gravity. //

TUTOR: Times acceleration due to gravity. //

Don't say gravity. //

On the other hand, a tutor can also give an *elaborative* feedback to an error, such as:

TUTEE: FN would be//

would FN be mass of A plus mass of B? Or? //

TUTOR: Again you- a force cannot be mass. //

These are two distinct quantities. //

Examining these latter three forms of feedback (negative, corrective, and elaborative) corresponds to analyzing interactive dialogue units of an error followed by a feedback segment.

Table 5 (1st column) shows the average number of feedback-to-error dialogue units for the Good and the Poor Tutees. It is not surprising that there are almost twice as many feedback-to-error units for the Poor Tutees since they committed more errors during tutoring ($M = 89$) than the Good Tutees who committed fewer errors ($M = 56$), as mentioned earlier. Given that errors tend to elicit feedback, the contrast in the frequency of feedback-to-errors between the Good and the Poor Tutees makes sense.

Insert Table 5

The contrastive approach clarifies a possible reason for the puzzling overall marginal negative correlation between the Tutor's feedback and the Tutees' learning in Table 3. When the Tutor's feedback-to-errors were correlated separately for the Good

and the Poor Tutees, the overall marginal, negative correlation became an even stronger, negative correlation for the Poor Tutees only ($r = -.882$, $p = .048$, see column 2, Table 5). This suggests that the detrimental effect of the Tutor's feedback only affected the Poor Tutees, whereas the Tutor's feedback-to-errors had no effect on the Good Tutees' learning.

A similar pattern of a strong negative correlation between the Tutor's feedback-to-errors and learning occurred for the Collaborative Observers as well ($r = -.835$, $p = .003$, see 3rd column in Table 5). That is, the Observers suffered when they observed the Poor Tapes, in terms of the frequency of feedback-to-errors.

The interpretation we offer is the following. Feedback-to-errors have no effect on the Good Tutees perhaps because they can learn even without feedback; that is, they can ignore the feedback. The Poor Tutees, on the other hand, could not benefit from the Tutor's feedback to their errors perhaps they could not make sense of the feedback, so that the more feedback they received to their errors, the more confused they were (thus less learning); and such confusion might have affected the learning of the Observers who watched their tapes. Recall that we reported earlier that the Poor Tutees overall did express confusion twice as frequently as the Good Tutees. Thus, the Poor Tutees had difficulty making sense of the Tutor's feedback.

But why might the Poor Tutees have difficulty making sense of the Tutor's feedback? One possible reason is that the feedback they received was less informative. For example, corrective feedback is less informative because it only gave the correct answer without further justifications, as in elaborative feedback. Table 5 (the last 3 columns) shows a distribution of the three different types of feedback (negative, corrective &

elaborative) to errors. Although the distribution of the three types of feedback the Tutor gave is similar for the Good versus the Poor Tutees (both groups received the lowest proportion of elaborative feedback and highest proportion of corrective feedback), Poor Tutees received proportionately more corrective feedback than elaborative feedback (52% vs 16%), as compared to Good Tutees (43% vs 23%). The contrast in the difference between the corrective and elaborative feedback for the Poor Tutees (36%) and the Good Tutees (20%) was significant ($F(1,8) = 5.188$, $p = .052$). In other words, the Poor Tutees received significantly more corrective feedback than elaborative feedback ($F(1,4) = 32.106$, $p = .005$), whereas there was no significant difference in the two types of feedback received by the Good Tutees. Because corrective feedback is less elaborative and contains no justifications, it may be more difficult for Poor Tutees to make sense of corrective feedback, which is the predominant kind of negative feedback that they received. Not making sense of the corrective feedback they received in turn effected how well the Observers could learn from overhearing them as well.

Thus, the earlier analyses, of looking only at Tutor's feedback as independent moves, masked much stronger correlational effects when we analyzed feedback in an interactive and contrastive way. Basically, a tutor's feedback-to-errors seems harmless to Good Tutees and their Observers, but detrimental to Poor Tutees and their Observers. This suggests that feedback per se is not the only critical factor, but what kind of feedback a tutor gives, and whether or not tutees can assimilate, understand, and use the feedback, thus supporting the *interaction hypothesis*. As reported earlier, the Poor Tutees expressed more confusion than the Good Tutees, perhaps because the feedback they

received was less elaborative. Overhearing the Poor Tutees' feedback-to-error dialogue units must have had a detrimental effect on the Observers' learning as well.

4.2.2.3 Summary of dialogue-level analyses. The first set of analyses showed that the most effective form of dialogue units are scaffoldings followed by relevant substantive Tutee responses (see last 2 columns in Table 4), in terms of both the Tutees and the Observers' learning. We surmise that Tutees learned from them because they could jointly construct meaningful follow-up responses to the Tutor's scaffoldings, but less so to the Tutor's explanations. We assumed that jointly constructed dialogues may be shorter and more coherent, and *scaffolding*-relevant response dialogue units did turn out to be more coherent than *explanation*-response dialogue units, based on the Coh-Metrix analysis. The finding that the Observers also learned only when they overheard *scaffolding*-relevant response dialogue units is consistent with the coherence interpretation. Additionally, *scaffolding*-response units may be more understandable than *explanation*-response units because they tend to be shorter, as confirmed by the word count analysis. These findings provide evidence in support of *the interaction hypothesis*.

The second set of analyses examined feedback-to-errors. We found that feedback-to-errors was detrimental to Poor Tutees and Observers of their tapes, but not to the Good Tutees and Observers of their tapes. The interpretation we offered was that Poor Tutees needed the feedback, and yet possibly could not make sense of the feedback since the Tutor's feedback to them were more of the corrective kind rather than the elaborative kind. Corrective feedback basically gave only the right answer whereas elaborative feedback gave the justification as well. In short, Tutees' learning is a function of both whether or not they can make sense of a tutor's feedback as well as whether a tutor gives

them more elaborative feedback, again, supporting *the interaction hypothesis*. Thus, the differential learning gains of the Good versus the Poor Tutees as a function of feedback-to-errors further underscore the importance of the role of the tutees, in being able to make sense of the feedback, and not merely the role of a tutor, in terms of whether the right kind of feedback was given or not.

4.2.3   Episode Analyses

In the prior dialogue analyses, we inferred that scaffolding-relevant follow-up units were jointly constructed because they were more coherent. However, we can directly code dialogue units as either jointly constructed or independently constructed by looking at a larger grain size. This would allow us to test the *interaction hypothesis* more directly. Accordingly, another pass at coding the protocols was undertaken at a larger episode-level grain size.

Segments in the tutoring protocols were combined into episodes. An episode is usually a multi-turn dialogue unit bounded by utterances whose content is relevant to a specific solution node (as shown in Fig. 1). Appendix B illustrates several episodes. For example, Episode III is relevant to Node 2.2.2 in Figure 1. The appendix in its entirety can be found at http://www.cogsci.rip.edu/CSJarchive/Supplemental/Index.html.

4.2.3.1. Joint and independent coverage of all nodes. For each episode, we differentiated whether the substantive contributions were initiated and covered by the Tutor alone (as in Episode II), the Tutees alone (as in Episode III), or jointly by both the Tutor and the Tutees (as in Episodes IV, V, see Appendix B).

Table 6 shows that a majority of the episodes (55 per tutoring session) were jointly covered by the Tutor and the Tutee, followed by 32 episodes covered

independently by the Tutor, and 16 independently covered by the Tutees. If we assume that joint coverage involves more scaffolding and independent Tutor coverage involves more explaining, then this difference between the frequency of joint coverage and independent Tutor coverage mirrors the results of greater frequency of Tutor scaffolding than explaining (see Table 3 again).

Insert Table 6 about here

If interacting with the Tutor facilitates learning, then there should be a significant correlation between the frequency of joint coverage and Tutees' learning. Table 6 shows that Tutees indeed learned when they jointly covered a node with the Tutor ($r = 0.646$, $p = .043$), thereby supporting the *interacting hypothesis*. Moreover, the Tutees also learned when they covered the nodes independently ($r = 0.637$, $p = .047$) suggesting that independent coverage obviously required them to be constructive, thereby leading to learning, thus supporting the *student-centered hypothesis*. The significant correlation of the Tutees' independent coverage of nodes replicate the significant correlation of Tutees' Substantive moves at the segment level (Table 3). Thus, analyses at two different grain sizes produce the same pattern of results. Finally, the Tutees did not learn when the Tutor independently covered a node (Table 6), just as they did not learn when the Tutor's scaffolding and explaining moves were considered independently (Table 3), thus weakening the important of the *tutor-centered hypothesis*.

Can the Observers' learning further confirm the *interacting* and the *student-centered* hypotheses as well as any of our prior interpretations? The Observers likewise

learned from overhearing joint coverage of nodes ($r = 0.457$, $p = .043$) but not from the Tutor's coverage of nodes. Again, if we assume that joint coverage involves more scaffolding and independent Tutor coverage involves more explaining, then the same interpretation offered for the results reported in Table 5 can be applied here as well: that is, Observers learn from overhearing joint coverage perhaps because joint coverage are more coherent and short, containing scaffolding-response dialogues, whereas Tutor's independent coverage may be less coherent and long. The Collaborative Observers also benefited somewhat from overhearing Tutees' independent coverage of a node (a trend). Their weaker learning from overhearing the Tutees' independent coverage of a node (as compared to joint coverage) reinforces the interpretation that a Tutees' construction often serves their own purposes, and may be less comprehensible to others (Chi, 2000), consistent with the lack of correlation between the Observers' learning and Tutees' irrelevant substantive responses (see Table 3 again).

The contrast between the Observers' learning from overhearing the Tutees' independent coverage but not from overhearing the Tutor's independent coverage, is related to some findings in the literature with respect to learning from an expert versus a peer. For example, Hinds, Patterson and Pfeffer (2001) have found that learners performed better when instructed by novices than by experts in an electronic wiring task. Likewise, Cho, Schunn, and Charney (2006) found that students are far more able to incorporate feedback from their peers than from their instructor in a writing task. These findings, along with the result here of both the Tutees' and the Observers' failure to learn from the Tutor's independent coverage, are consistent with the finding that Tutees also do

not learn from receiving the Tutor's explanations (as shown by a lack of correlation in Table 3).

4.2.3.2 Summary of Episode Analyses. In sum, analyses at the episode level provide further evidence to support both *the student-centered* and the *interaction hypothesis,* but no evidence in support of the *tutor-centered hypothesis.* The correlation of learning with joint coverage of problem solving nodes supported the *interaction hypothesis*, while the correlation of learning with the tutee's independent coverage of the nodes supported the *student-centered hypothesis.* The lack of any correlation of learning with Tutor's independent coverage further undermines the *tutor-centered hypothesis*.

4.3 Is the Tutor Adaptive?

We proposed above that the *tutor-centered hypothesis* can be evaluated in terms of the frequency and quality of a tutor's moves, and a tutor's adaptiveness. Adaptiveness includes choice of moves, timing of moves, and assessment of tutees' understanding. Assessment furthermore can take one of two forms. Normative assessment means a tutor evaluates whether a tutee is getting something right, as a function of the formal knowledge of a discipline. There is no question that tutors, knowledgeable about their domains, are competent at doing this, and base their feedback on normative knowledge (Putnam, 1987). However, assessment can also mean evaluating a tutee's understanding from the students' perspective, such as knowing what a tutee's misunderstanding is, or in what ways is a tutee's mental model flawed. Our prior findings have shown that neither an expert tutor nor novice tutors were very accurate at assessing their tutees' understanding. In the domain of physics, we found an expert tutor to often overlook the tutee's misunderstanding (Chi, 1996). In the domain of the circulatory system, when we

directly assessed tutors' conceptions of their tutees' mental model of the circulatory system while tutoring, we found their conceptions to be completely inaccurate (Chi et al., 2004). If a tutor cannot accurately assess a tutee's misunderstanding from a student's perspective, then they cannot be adaptive from this student-model perspective.

In addition, the main finding reported in this paper, the fact that the Observers can learn as well as the Tutees, also suggests indirectly that adaptiveness cannot be a powerful feature of tutoring effectiveness, since the Tutor obviously could not be adapting to the specific needs of the Observers. A specific finding of this study also questions the adaptiveness of the Tutor. This is the result showing that the Tutor gave more corrective than elaborative feedback to the Poor Tutees (Table 5). One could argue that this is counterproductive and not adaptive in that the Poor Tutees needed more elaborative feedback. Aside from these negative findings, can we provide any evidence for a tutor's adaptiveness?

One manifestation of adaptiveness, as suggested above, is to know what kind of problems to present to the tutees. For example, one would expect that the Tutor would choose the more challenging problems for the Good Tutees. As noted in Appendix A, the Tutor designed the four tutoring problems and rank ordered their difficulty from easiest (Problem 1) to the hardest (Problem 4). Therefore, a simple but gross measure of his adaptiveness is to see whether or not he presented the more challenging problems to the Good Tutees. Although the hardest Problem 4 was presented only twice to the Good Tutees, to get more data points, we combined the two easier problems (1 and 2) and the two harder problems (3 and 4), and see which ones were assigned to be tutored. There was absolutely no difference between the Good and the Poor Tutees: On average, the Tutees in

both groups were assigned the easier problems exactly 7 times and the harder problems exactly 3 times. Thus, although this measure of adaptiveness (using only 2 sets of problems) is not very sensitive, the result at minimum suggests that we cannot conclude that the Tutor was adaptive in terms of his selection of which problems to use in tutoring, as a way to tailor his instruction to the competence and needs of a tutee.

5. Conclusion

The two overarching goals of this research were to design an alternative learning environment that facilitates learning from tutoring and can be more easily adapted and scaled up for a variety of learning contexts, such as on-line as well as classroom learning; and to further our understanding of tutoring effectiveness. The alternative learning environment is observing tutoring collaboratively. The motivation for designing such an environment was to leverage and combine the benefits of tutoring, collaborating, and the potential benefit of observing.

Comparing five different contexts of learning (tutoring, observing collaboratively, collaborating, observing alone, and studying alone), we replicated the tutoring literature by showing that human tutoring is indeed the most effective form of instruction. More interestingly, students in the alternative learning environment of observing tutoring collaboratively, seemed to learn as effectively as students participating in tutoring. In fact, when one observes tutoring of competent tutees (the Good Tutees), then learning by the Observers is totally equivalent to (in fact slightly exceeded) learning by the Tutees who participated interactively in the tutoring. Such results suggest that it is interacting *per se* that may account for learning (such as interacting with a peer), and not necessarily interacting with a tutor. Even more impressive are the learning gains of Collaborative

Observers who were poor learners, when they watched good learners being tutored. Since learning from tutoring has historically manifested the highest amount of learning gains, to find another learning environment that can match and exceed this benchmark, in the case of the poor learners, is extremely promising.

The results from the target Observing Collaboratively condition provided one explanation for the discrepancy in the literature about the benefit of observational learning for complex cognitive tasks. The conjecture we proposed is that the discrepancy in the literature may have arisen from the variability from study-to-study in terms of how active and engaged the observer was. This *active/constructive/interactive observing* hypothesis was supported in four ways in this study. First, the mere fact that the Collaborative Observers could learn as well as the Tutees who participated in tutoring, suggests that this can be accounted for by the interactions of the Collaborative Observers per se, without interacting directly with a tutor. Second, the fact that the more interactive the Collaborators Observers were with each other while observing, the more they learned, lends further support to the *active/constructive/interactive observing hypothesis*. Third, two conditions of observations were manipulated, observing collaboratively and observing alone. Collaborative Observers learned more than the Lone Observers, confirming our hypothesis that being more constructive, which was more likely in the collaborative condition, will enhance learning from observing. Finally, we teased apart the Lone Observers into those who were active versus those who were more passive. Again, we found that the active Lone Observers learned more than the more passive Lone Observers, again confirming our hypothesis. In short, these four results seem to support unambiguously the conjecture that variability in engagement might have explained the

discrepancy in the literature, especially given that the task predominantly used in the vicarious learning literature involved a single observer only.

The research findings shed further light on explaining why tutoring is effective. Consistent with our prior results (Chi et al., 2001), there is no question that Tutees learned when they made substantive contributions, regardless of the relevancy of their substantive contributions (see Table 3), as well as when they covered nodes themselves (Table 6), thus supporting the *student-centered constructive hypothesis.*

The Tutor can contribute toward Tutees' learning more by scaffolding the Tutees than by explaining to them (see Table 4), presumably because scaffolding enabled the Tutees to jointly construct coherent follow-up responses, thus supporting the *interactive coordination hypothesis.* The Tutees also learned when they covered a node jointly with the Tutor. Finally, the Tutor's feedback-to-errors were particularly harmful to the Poor Tutees and their Observers, consistent with the interpretation that Poor Tutees found the feedback confusing, in part because the Tutor gave them more corrective than elaborative feedback. Thus, it is not the feedback *per se* that determined the Tutees' learning, but whether the Tutees could make sense of the feedback, and what kind of feedback the Tutor were more likely to give. In short, there is substantial evidence to support the *interaction hypothesis*, revealed largely from analyzing the protocol data at a both a two-turn dialogue unit and multi-turn episode unit of analyses.

Although tutoring effectiveness has traditionally been attributed to the skill of the tutor *per se*, we could find no evidence of learning from considering the tutor's contributions alone, using analyses at various grain sizes. Although some tutor moves were better than others (Scaffolding is better than Explaining, see Table 4), they were

better because of the responses the Tutees could construct, and not because they were good instructional moves themselves (see Table 3). Such results extend and support our previous findings and conjectures (Chi et al., 2001), wherein we could only show in an indirect way that tutors' explanations were not effective by suppressing tutors' explanations. However, given that those were novice tutors, one may be skeptical of the results of that prior study because those tutors' explanations may have been of lower quality (although such skepticism is not well-founded given the results of by Cho et al., 2006 and Hinds et al., 2001, showing that learners benefit more from novice instructors). The results of the current study, using a more expert-like tutor who presumably was capable of giving quality explanations, re-confirm the interpretation of the prior study. Overall, there was very little evidence to support the *tutor-centered hypothesis* in terms of frequency, quality, and adaptiveness (albeit we have scant evidence for the adaptiveness component of this hypothesis). In short, in a very different domain, with a different age group, these data once again undermine the influence of the tutor, and further support the contributions of the tutees themselves, and their interactions with the tutor, as responsible for learning.

Although several of our analyses relied extensively on correlations, we believe that in the majority of the cases, we can infer causality since the intervention occurred prior to the learning outcome. Moreover, we relied on the patterns of correlations for our interpretations rather than on the values of individual correlations. Finally, despite a low sample size, the correlations were often quite substantial and learning was measured in terms of long-term (several days post-training) deep knowledge gains.

The finding that students can learn by observing tutoring vicariously has promising implications for how individual classroom interactions between a teacher and a student (such as at the blackboard) can be productive for the rest of the students in the class. Our finding basically supports Rogoff and colleagues' argument that learning through observation is a valuable but often overlooked practice in mainstream schooling (Rogoff, Paradise, Arauz, Correa-Chavez, & Angelillo, 2003). Such finding also has implication for how databases of re-usable dialogues can be created as a resource for distance learners (Stenning et al., 1999) and for on-line environments (Hudson & Bruckman, 2004). Finally, the findings that observers tend to learn in dialogues that involve the Tutees emphasize the importance of learning from peers, from both teaching a peer (Roscoe & Chi, under review, a) as well as receiving comments and questions from a peer (Roscoe & Chi, under review, b).

References

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive

tutors: lessons learned. *The Journal of the Learning Sciences, 4*, 167–207

Anzai, Y. & Simon, H.A. (1979). The theory of learning by doing. *Psychological*

Review, 86, 124-140.

Bandura, A. (1969). *Principles of behavior modification.* New York, NY: Holt, Rinehar,

& Winston.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory.*

Englewood Cliffs, NJ: Prentice Hall.

Biswas, G., Schwartz, D., Bransford, J., & the Teachable Agent Group at Vanderbilt

(TAG-V) (2001). Technology support for complex problem solving: From SAD

environments to AI. In K.D. Forbus and P.J. Feltovich (Eds.), *Smart machines in*

education (pp. 71-98). Menlo Park, CA: AAAI Press.

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.

Chi, M.T.H. (1996). Constructing self-explanations and scaffolded explanations in

tutoring. *Applied Cognitive Psychology, 10*, 33-49.

Chi, M. T. H. (1997). Quantifying qualitative analyses of verbal data: a practical guide.

Journal of the Learning Sciences, 6, 271–315.

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., 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. & Bjork, R. A. (1991). Modeling expertise. In Druckman & R. A. Bjork (Eds.), *In the Mind's Eye: Enhancing human performance* (pp. 57-79). Washington, D.C.: National Academy Press.

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., McGregor, M.U., & Hausmann, R.G. (2000). Learning from overheard tutorial dialogue: Benefits of self-explanation. Paper presented at the Third Annual Highlands Undergraduate Psychology Conference, University of Pittsburgh at Johnstown, April.

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

Chi, M. T. H., Siler, S., & Jeong, H. (2004). Can tutors diagnose student's misunderstandings? *Cognition and Instruction*, *22*, 363-387.

Chi, M. T. H., Siler, S., Jeong, H., Yamauchi, T., & Hausmann, R. (2001). Learning from human tutoring. *Cognitive Science*, *25*, 471-534.

Cho, K.., Schunn, C.D., & Charney, D. (2006). Commenting on writing: Typology and perceived helpfulness of comments from novice peer reviewers and subject matter experts. *Written Communication,* 23, 260-294.

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.

Collins, A., & Stevens, A.L. (1982). Goals and strategies of inquiry teachers. In R. Glaser (Ed.), *Advances in instructional psychology (Vol. 2).* Hillsdale, NJ: Erlbaum.

Collins, A., Brown, J.S. & Holum, A. (1991) Cognitive apprenticeship: Making thinking visible. *American Educator*, Winter, 6-46.

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 453-494). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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

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

Craig, S. D., Sullins, J., Witherspoon, A. & Gholson, B. (2006). Deep-Level Reasoning

Questions effect: The Role of Dialog and Deep-Level Reasoning Questions

during Vicarious Learning. *Cognition and Instruction, 24(4),*563-589.

Driscoll, D., Craig, S.D., Gholson, B., Ventura, M., Hu, X., & Graesser, A. (2003).

Vicarious learning: Effects of overhearing dialogue and monolog-like discourse in

a virtual tutoring session. *Journal of Educational Computing Research, 29*, 431-

450.

Eisenhart, M., Borko, H., Underhill, R., Brown, D., Jones, D., & Agard, P. (1993).

Conceptual knowledge falls through the cracks: Complexities of learning to teach

mathematics for understanding. *Journal for Research in Mathematics Education,

24*, 8-40.

Evens, M. W., Spitkovsky, J., Boyle, P., Michael, J. A., & Rovick, A. A. (1993).

Synthesizing tutorial dialogues. In *Proceedings of the Fifteenth Annual

Conference of the Cognitive Science Society*, 137–140. Hillsdale, New Jersey:

Lawrence Erlbaum Associates.

Fox Tree, J.E. (1999). Listening in monologs and dialogues. *Discourse Processes, 27*,

35-53.

Frederiksen, C.H., Donin, J., Meilleur, L., Roy, M., & Bracewell, B. (1999). *Learning

through tutorial dialogue in a computer-based problem-solving environment:

Bystander learning*. Paper presented at annual meeting of the American

Educational Research Association, Montréal, April.

Graesser, A. C., & Person, N. (1994). Question asking during tutoring. *American

Educational Research Journal*, *31*, 104-137.

Graesser, A. C., Person, N., & Magliano, J. (1995). Collaborative dialog patterns in

naturalistic one-on-one tutoring. *Applied Cognitive Psychology, 9,* 359–387.

Halliday, D. & Resnick, R. (1981). *Fundamentals of physics* (2nd ed.). New York, NY:

Wiley.

Hinds, P.J., Patterson, M. & Pfeffer, J. (2001). Bothered by abstraction: The effect of

expertise on knowledge transfer and subsequent novice performance. *Journal of

Applied Psychology,* 86, 1232-1243.

Hudson, J. M., & Bruckman, A.S. (2004). The bystander effect: A lens for understanding

patterns of participation. *The Journal of the Learning Sciences, 13*, 165-195.

Hume, G. D., Michael, J. A., Rovick, A. A., & Evens, M. W. (1993). The use of hints as

a tutorial tactic. In M. C. Polson (Ed.), *Proceedings of the fifteenth annual

conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum

Associates.

Jackson, G., Person, N. and Graesser, A. (2004). Adaptive tutorial dialogue in

AutoTutor. In *Proc. Workshop on dialog-based Intelligent tutoring Systems at

Intelligent Tutoring systems*.

Johnson, D. W., & Johnson, R. T. (1992). Positive interdependence: Key to effective

cooperation. In R. Hertz-Lazarowitz & N. Miller (Eds.), *Interaction in

cooperative groups: The theoretical anatomy of group learning*, (pp. 174-199).

Cambridge, England: Cambridge University Press.

King, A. (1992). Comparison of self-questioning, summarizing, and notetaking-review as

strategies for learning from lectures. *American Educational Research Journal*, *29*,

303-323.

Latham, G. P. , & Saari, L. M. (1979). Application of social-learning theory on training

  supervisors through behavior modeling. *Journal of Applied Psychology, 64*, 239-

  246.

Lepper, M. R., Woolverton, M., Mumme, D. L., & Gurtner, J. L. (1991). Motivational

  techniques of expert human tutors: Lessons for the design of computer-based

  tutors. In S. P. Lajoie & S. J. Derry (Eds.), *Computers as cognitive tools*, (pp. 75-

  105). Hillsdale, NJ: Lawrence Erlbaum Associates.

Litman, D., & Forbes-Riley, K. (2006). Correlations between dialogue acts and learning

  in spoken tutorial dialogues. *Natural Language Engineering, 12*, 161–176.

McKendree, J., Good, & J., Lee, J. (2001). Effects of dialogue characteristics on

  performance by overhearers. *Cognitive Science Society Meetings*, Edinborough,

  Scotland.

McKendree, J., Stenning, K., Mayes, T., Lee, J., & Cox, R. (1998). Why observing a

  dialogue may benefit learning: The Vicarious Learner. *Journal of Computer

  Assisted Learning, 14,* 110-119.

Merrill, D. C., Reiser, B. J., Merrill, S. K., & Landes, S. (1995). Tutoring: guided

  learning by doing. *Cognition and Instruction, 13*, 315–372.

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*, 277–306.

McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes,

  38*(1), 1-30.

McNamara, D. S., Louwerse, M. M. Cai., Z., & Graesser, A. (2005, January 1) Coh-

Metrix version 1.4. Retrieved ,06/1/06 from http//:cohmetrix.memphis.edu.

Meltzoff, A. N. (2005). Imitation and other minds: The "like me" hypothesis. In S.

Hurley and N. Chater (Eds.), *Perspectives on imitation: From neuroscience to*

social science. (Vol. 2, pp. 55-77). Cambridge, MA: MIT Press.

Murray, R. C. & VanLehn, K. (2006). A comparison of decision-theoretic, fixed-policy

and random tutorial action selection. In M. Ikeda, K. Ashley, and T.-W. Chan

(Eds.): ITS 2006, LNCS 4053, (pp. 114-123). Berlin: Springer-Verlag.

Nisbett, R.E. & Wilson, T.D. (1977). Telling more than we can know: verbal

reports on mental processes. *Psychological Review, 92*, 289-316.

Okita, S. Y., Schwartz, D. L. (2006). When observation beats doing: Learning

by Teaching, In S. Barab, K. Hay, D. Hickey (Eds.), 7th International Conference

of the Learning Sciences: Vol. 1 (pp.509-516). New Jersey: Lawrence Erlbaum

Associates.

Pilkington, R. & Parker-Jones, C. (1996). Interacting with computer-based simulation:

The role of dialogue. *Computers and Education*, 27, 1-14.

Putnam, R. T. (1987). Structuring and adjusting content for students: A live and

simulated tutoring of addition. *American Educational Research Journal*, *24*, 13-

48.

Roscoe, R. & Chi, M. (Under review, a) Reflective knowledge-building and knowledge-
telling in peer tutors' explanations and questions. *Review of Educational
Research.*

Roscoe, R. & Chi, M. (Under review, b) Tutor learning: The role of explaining and
responding to questions. *Instructional Science*.

Rogoff, B., Paradise, R., Arauz, R. M. , Correa-Chavez, M., & Angelillo, C. (2003).

    Firsthand learning though intent participation. *Annual Review of Psychology*, *54*,

    175-203.

Rummel, N. & Spada, H. (2005). Learning to collaborate: An instructional approach to

    promoting collaborative problem solving in computer-mediated settings. *The*

    Journal of the Learning Sciences, 14, 201-241.

Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and observers.

    Cognitive Psychology, 21, 11–232.

Schwartz, D.L., Blair, K.P., Biswas, G., Leelawong, K, & Davis, J. (In press). Animation

    of thought: Interactivity in the teachable agent paradigm. In R. Lowe & W.

    Schnotz (Eds.), *Learning with Animation: Research and Implications for Design*.

    UK: Cambridge University Press.

Shulman, L.S. (1986). Paradigms and research programs in the study of teaching: A

    contemporary perspective. In M.C. Wittrock (Ed.), *Handbook of Research on*

    Teaching (3rd ed., pp. 3-36). New York: Macmillan.

Slavin, R. E. (1990). *Cooperative learning: Theory, research, and practice*. Englewood

    Cliffs, NJ: Prentice Hall.

Sleeman, D., Kelly, A. E., Martinak, R., Ward, R. D., & Moore, J. L. (1989). Studies of

    diagnosis and remediation with high school algebra students. *Cognitive Science*

    13, 551-568.

Stenning, K., McKendree, J., Lee, J., Cox, R., Dineen, F., Mayes, T. (1999). Vicarious

    learning from educational dialogue. Proceedings of *Computer-Supported*

    Cooperative Learning (CSCL '99),(pp. 341-347). Palo Alto: Stanford University.

Sweller, J., van Merrienboer, J.J.G., & Paas, F.G. (1998). Cognitive architecture and

instructional design. *Educational Psychology Review,* 10, 251-296.

Tao, P.K., & Gunstone, R.F. (1999). Conceptual change in science through collaborative

learning at the computer. *International Journal of Science Education,* 21, 39-57.

Tomasello, M. (1999). *The cultural originas of human cognition.* Cambridge, MA:

Harvard University Press.

Traum, D., & Heeman, P. (1997). Utterance Units in Spoken Dialogue. In E. Maier, M.

Mast, & S. LuperFoy (Eds.), *Dialogue Processing in Spoken Language Systems,*

(pp. 125-140). Heidelberg: Springer-Verlag.

VanLehn, K., Siler, S., Murray, C., Yamauchi, T., & Bagget, W. B. (2003). Human

tutoring: Why do only some events cause learning? *Cognition and Instruction*, *21*,

209-249.

VanLehn, K., Lynch, C., Schulze, K., Shapiro, J. A., Shelby, R. H., Taylor, L., Treacy, D.

J., Weinstein, A., & Wintersgill, M. C. (2005).   The Andes physics tutoring

system: Five years of evaluations. In G. I. McCalla and C.-K. Looi (Eds.),

Proceedings of the Artificial Intelligence in Education Conference. Amsterdam:

IOS.

Webb, N. M. (1989). Peer interaction and learning in small groups. In Webb, N. (Ed.)

Peer interaction, problem solving, and cognition: Multidisciplinary perspectives.

[Special issue]. *International Journal of Education Research, 13*, 21-39.

Wood, D. J., Wood, H., & Middleton, D. (1978). An experimental evaluation of four

face-to-face teaching strategies. International Journal of Behavioral Development,

1, 131-147.

Appendix A

Pre-test, Tutoring, and Post-test Problems.

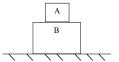| PRE TEST PROBLEM	TUTORING PROBLEM	POST TEST PROBLEM

Warm-up Problem

A mass m is acted on by two forces F_1 and F_2, as shown. If $m = 5.0$ kg, $F_1 = 5.0$ N and $F_2 = 6.0$N, find the vector acceleration of the body.

Warm-up Problem

Block B is resting on the ground. Block A is sitting on B. Masses of A and B are M_A and M_B.
a) Draw free body diagrams for A, B, and G.
b) Write equations of motion for A and B.
c) Group action reaction pairs for the forces labeled above.

Tutor Rationale: This is simplest of the tutoring problems involving a static situation. The tutor used this first with all students to assess and shore up their basic understanding of Newton's three laws of motion. According to the tutor, the most common problems expected from students on this problem include difficulty identifying the forces that act on a body, understanding net force, and understanding that action reaction pairs act on separate entities.

Problem 1

A block m = 30 kg is attached to a string, which is tied to the wall. The block is resting on a smooth incline at an angle θ = 40° with the floor. What is the tension in the string?

Bock A is attached to a string that is tied to a wall. The block is resting on a smooth plane inclined at an angle θ with the horizontal.
a) What is the tension in the string?
b) What is the force exerted by the wall on the string?
c) The string is now cut. Find the acceleration of the block if θ = 30° and MA = 10 kg.

Tutor Rationale: This problem introduces motion, and requires decomposition of the weight force into components. The tutor expected student difficulty dealing with components and with choosing an efficient coordinate system.

A block M = 20 kg is attached to a string, which is tied to the wall. The block is resting on a smooth incline at θ = 40° with the floor.
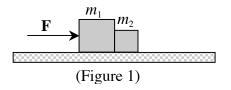a) What is the tension in the string?
b) What is the force exerted by the wall on the string?
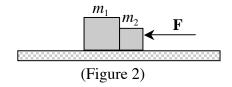c) Imagine that the string is cut. Find the acceleration of the block.

Problem 2

Two blocks are in contact on a frictionless table. A horizontal force is applied to one block.

a) If $m1 = 2.0$ kg, $m2 = 1.0$ kg, and $F = 3.0$ N, find the force between the two blocks.
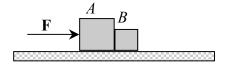

(Figure 1)

b) Show that if the same force $F = 3.0$ N is applied to $m2$ rather than $m1$ (see figure 2), the force of contact between the blocks is different than in Problem a. Explain why it is different.

(Figure 2)

Two blocks A and B are in contact with each other on a smooth floor. A force of 10N is applied to the blocks as shown in the figure. Masses of the blocks are 2 kg and 3 kg respectively.
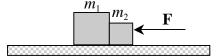a) Find the acceleration of the blocks
b) Find net force acting on block B.
c) Find force exerted by block B on A.

Tutor Rationale: The third problem involves two blocks of differing mass in contact with one another. Both blocks are sitting flat on the ground and a horizontal force is applied to one block. The trick with this problem is to treat the two blocks as a single compound body.

Two blocks are in contact on a frictionless table. A horizontal force is applied to one block, as shown below. If $m_1 = 5$ kg, $m_2 = 4$ kg, and $F = 20$ N:
a) Find the acceleration of the blocks
b) Find the net force acting on block m_2.
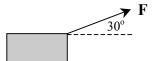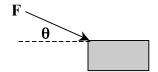c) Find the force exerted by block m_1 on m_2.

Problem 3

A person pulls a loaded wagon having a total mass of 70 kg (see figure below). A force of 100 N applied through the handle at 30°. Ignoring friction, what is the resulting horizontal acceleration?

A person pushes a crate on a smooth floor. He is applying force at an angle θ with the horizontal.
a) Find the normal reaction of the floor on the crate.
b) If the mass of the crate is 10 kg, the magnitude of the force is 5N and θ = 30° what will be the acceleration of the crate?

A person pushes a 20 kg block at an angle of θ = 30° with the floor (see figure below). A force of 20 N is applied to the block. Ignoring friction:
a) Draw the free body diagram for the block
b) Find the normal reaction of the floor on the block.
c) What is the acceleration of the block?
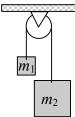
F
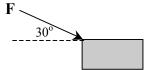
30°

F

θ

F

30°

Tutor Rationale: The fourth problem involves a block sitting flat on the floor with a downward force applied at an angle. The difficulty with this problem is that involves decomposing forces into components and also requires students to understand that normal force can be affected by an applied force.

Problem 4

Two blocks, m_1 and m_2 are connected to two ends of a weightless string which passes over a weightless frictionless pulley. The pulley is supported from the ceiling.
a) Find acceleration of the blocks.
b) Find tension in the string.
c) What is the force exerted by the pulley support on the ceiling?
d) If $m1 = 4.2$ kg, $m2 = 5.6$ kg, how far will $m2$ be displaced in 4s?
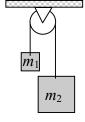
Two blocks, $m_1 = 10$ kg and $m_2 = 20$ kg, are connected by a string that passes over a frictionless and massless pulley.
a) Find the tension in the string
b) Find the acceleration of the masses.

Tutor Rationale: The fifth Tutoring problem is the most challenging and involves two blocks of different mass suspended by a string supported by a pulley. The complexity of this problem is that it involves two bodies in motion.

Appendix B

Example of Episode Coding.

TUTOR: Now umm in this problem we have now two blocks.
Again two blocks.
But they are sitting one against the other.   Episode I Node 2.1
This is A and this is B.
There is a force F acting on it.
Okay.

TUTEE: ummhumm.

TUTOR: Now, what will be the acceleration of these blocks?
Umm you can start by assuming the mass of block A is MA   Episode II Node 2.2
and mass of block B is MB.

TUTEE: Umhm.
Acceleration.
[pause] I think I'm gonna be totally wrong.
But I'll try and do it. [S writes $\mathbf{F} = M*\mathbf{a}$; $\mathbf{a} = \mathbf{F}/M$]   Episode III Node 2.2.2
So. [pause]
The acceleration of both blocks?

TUTOR: Will they move together?

TUTEE: By- if that force pushes on them then [pause]
then [pause]
Ok well [laughs] /basically/

TUTOR: / (Hmmmm)/

TUTEE: /I guess it would./

TUTOR: /(In a)/ situation you have one box.   Episode IV Node 2.2
and you have another box against you then you push it.

TUTEE: /Ya I see./

TUTOR: /Will they/ move together? [laughs]

TUTEE: Well ya they would. But

TUTOR: /Ok./

TUTEE: /There/ are times when if you pushed it like
really fast then one would go and the other one would stay.
Kind of like those blocks that (hit) together.

TUTOR: Hitting is different.
You are impact

TUTEE: Oh Ok. [laughs]

TUTOR: Hitting is is an impact.
We are not talking of that /ah/impact

TUTEE: /Alright./

TUTOR: situation yet.
But if you have two blocks and you push it.

TUTEE: They both move together.

TUTOR: They will move together.
So what will be they're acceleration?

Episode V Node 2.1

TUTEE: So their acceleration will be force over
[long pause] [S writes $\mathbf{a} = \mathbf{F}/MA + MB$]
Is that totally wrong?

TUTOR: No why should it be totally wrong?
It's totally right.

TUTEE: Ok. [laughs]

TUTOR: Hmm.
This is (directly) Ok.

TUTEE: Ok. [laughs]

Episode VI
Node 2.2.2;
Node 2.2.2.1

TUTOR: Now ummm [pause] alright so- ah there are some
numbers given here.
You can use those numbers.

TUTEE: Alright.

TUTOR: Umm MA is given to be umm 2 Kilograms.

Episode VII
Node 2.2.2.1

MB is 3 Kilograms.
And force is 10 Newtons.

<div align="right">Episode VIII
Node 2.2.2.2</div>

TUTEE: MB equals 3 Kilograms? [S instantiates MA
and MB changing part of equation to "$a = F/2+3$"]

TUTOR: Yes.

Episode VII
Node 2.2.2.1 cont.

TUTEE: F equals what?

TUTOR: Ten Newtons. [pause]

Episode IX
Node 2.2.2.2

TUTEE: I can do this math.
[S writes $F = 10N$]

TUTEE: [pause] [S complete equation "$a = 10/2+3$']
S writes "$a = 2$ N/Kg"]

TUTOR: What is the name for Newtons per kilogram?

TUTEE: Oh oh.

TUTOR: What are the units of acceleration?

TUTEE: Acceleration is [long pause]
I don't know I don't have it.

TUTOR: Hmm Newtons per kilogram.

TUTEE: I have no idea.
Newtons per kilogram.

TUTOR: What is unit?

TUTEE: I would always- /I wo/

TUTOR: /Think/ in terms of acceleration.
What is the unit?

TUTEE: Meters per second squared.

TUTOR: Meters per second squared.

TUTEE: [S erases "N/Kg"]

TUTOR: So that is meters /per second squared./

TUTEE: /Oh that's it?/ [S writes "m/s^2"]

TUTOR: Mmmmm.

Episode X
Node 2.2.2.3

Table 1

The Number of Matched and Unique Shallow and Deep Steps

	Shallow Steps	Deep Steps	Total
Matched on Pre- and Post-test	6	16	22
Unique to Pre-test	2	3	5
Unique to Post-test	3	6	9

Table 2

Learning Resources and Learning Activities of the Five Conditions

	No. of Participants	Learning Resource	Learning Activity
Tutoring	$n = 10$	Human tutor	Interacting in dialogue Tutoring
Observing Collaboratively	$n = 20$	Peer Videotape	Interacting in dialogue Observing
Collaborating	$n = 20$	Peer Text	Interacting in dialogue Reading
Observing Alone	$n = 10$	Videotape	Observing
Studying Alone	$n = 10$	Text	Reading

Table 3

Correlation of Tutor and Tutee Moves with Tutees' and Collaborative Observers' Deep Learning

	Average No. of Segments per Session	Proportion	Tutees' Learning ($n = 10$)	Observers' Learning ($n = 20$)
Tutor Instructional Moves				
Scaffolding	245	36%	N.S.	N.S.
Explanation	157	23%	N.S.	N.S.
Feedback	130	19%	Trend $r = -.603, p = .065$	N.S.
Other	154	22%		
Total	686	100%		
Tutee Learning Moves				
Substantive	230	52%	Trend $r = .605, p = .064$	N.S.
Non-substantive	213	48%	$r = -.899, p = .000$	N.S.
Total	443	100%		
Tutee Substantive Moves				
Relevant follow-ups	99	43%	$r = .641, p = .047$	Trend $r = .398, p = .082$
Irrelevant responses	131	57%	Trend $r = .620, p = .056$	N.S.
Total	230	100%		

Table 4

Correlation of the Number of Tutor-Tutee Relevant Substantive Interactive Dialogue with Tutees' and Observers' Deep Learning

	Average No. per Session	Proportion	Tutees' Learning ($n = 10$)	Observers' Learning ($n = 20$)
Tutor scaffolding followed by Tutees' relevant substantive responses	59	60%	$r = .656, p = .039$	$r = .434, p = .056$
Tutor explaining followed by Tutees' relevant substantive responses	40	40%	Trend $r = .576, p = .082$	N.S.
Total	99	100%		

Table 5

The Average Frequency per Session, Correlation, and Distribution of Tutor Feedback to Good and Poor Tutees' Errors

				Breakdown of Feedback to Errors		
	Average No. of Tutor Feedback to Errors	Correlation of Average No. of Feedback with Tutee's Learning ($n = 10$)	Correlation of Average No. of Feedback with Observers' Learning ($n = 20$)	Negative Feedback	Corrective Feedback	Elaborative Feedback
Good Tutees	43	N.S.	N.S.	15 (34%)	19 (43%)	10 (23%)
Poor Tutees	80	$r = -.882,$ $p = .048$	$r = -.835,$ $p = .003$	26 (32%)	41 (52%)	13 (16%)
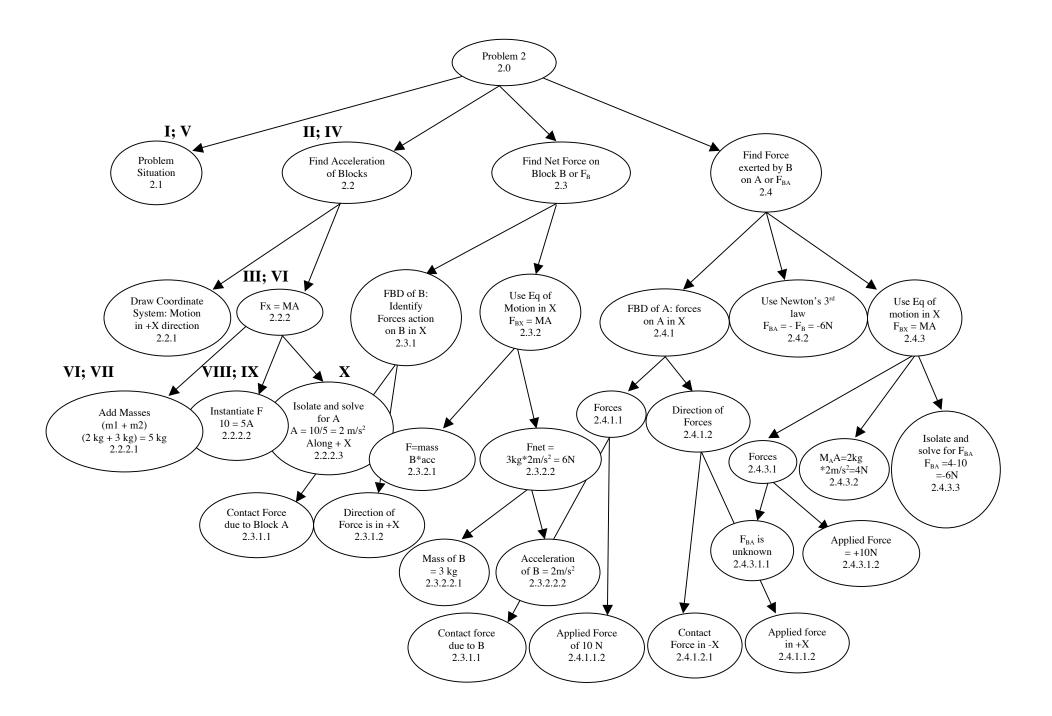
Table 6

Frequency and Correlations for All Node Episodes with Tutees' and Observers' Deep Learning
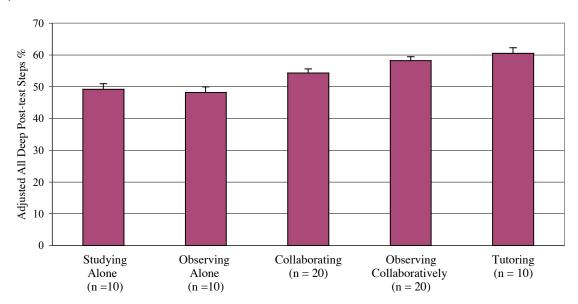
	Average Number Per Session	Tutees' Learning ($n = 10$)	Observers' Learning ($n = 20$)
Tutor	32	N.S.	N.S.
Tutor and Tutee	55	$r = 0.646, p = .043$	$r = 0.457, p = .043$
Tutees	16	$r = 0.637, p = .047$	Trend $r = 0.418, p = .067$

Figure Captions

Figure 1. Model of problem-solving nodes for the compound body

problem (Problem 2)

Figure 2a. Adjusted mean proportion of all deep post-test steps,

controlling for all deep pre-test steps for each condition

Figure 2b. Mean proportion of pre-test and post-test scores for matched

deep steps for each condition

Figure 3. Mean proportion of gain for matched deep steps for Good and Poor

Collaborative Observers watching Good and Poor Tutee tapes

Problem 2 2.0

I; V
Problem Situation 2.1

II; IV
Find Acceleration of Blocks 2.2

Find Net Force on Block B or F_B 2.3

Find Force exerted by B on A or F_{BA} 2.4

III; VI
Draw Coordinate System: Motion in +X direction 2.2.1

Fx = MA 2.2.2

FBD of B: Identify Forces action on B in X 2.3.1

Use Eq of Motion in X $F_{BX} = MA$ 2.3.2

FBD of A: forces on A in X 2.4.1

Use Newton's 3rd law $F_{BA} = - F_B = -6N$ 2.4.2

Use Eq of motion in X $F_{BX} = MA$ 2.4.3

VI; VII
Add Masses (m1 + m2) (2 kg + 3 kg) = 5 kg 2.2.2.1

VIII; IX
Instantiate F 10 = 5A 2.2.2.2

X
Isolate and solve for A A = 10/5 = 2 m/s^2 Along + X 2.2.2.3

F=mass B*acc 2.3.2.1

Fnet = 3kg*2m/s^2 = 6N 2.3.2.2

Forces 2.4.1.1

Direction of Forces 2.4.1.2

Forces 2.4.3.1

M_AA=2kg *2m/s^2=4N 2.4.3.2

Isolate and solve for F$_{BA}$ F_{BA} =4-10 =-6N 2.4.3.3

Contact Force due to Block A 2.3.1.1

Direction of Force is in +X 2.3.1.2

Mass of B = 3 kg 2.3.2.2.1

Acceleration of B = 2m/s^2 2.3.2.2.2

F_{BA} is unknown 2.4.3.1.1

Applied Force = +10N 2.4.3.1.2

Contact force due to B 2.3.1.1

Applied Force of 10 N 2.4.1.1.2

Contact Force in -X 2.4.1.2.1

Applied force in +X 2.4.1.1.2

a)

b)