Toward Intelligent Instructional Handoffs Between Humans and Machines

Steve Ritter
Carnegie Learning, Inc.
501 Grant St., Suite 1075
Pittsburgh, PA 15219-4447
sritter@carnegielearning.com

Stephen E. Fancsali
Carnegie Learning, Inc.
501 Grant St., Suite 1075
Pittsburgh, PA 15219-4447
sfancsali@carnegielearning.com

Michael Yudelson
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213
yudelson@cs.cmu.edu

Vasile Rus
University of Memphis
323 Dunn Hall
Memphis, TN 38152
vrus@memphus.edu

Susan Berman
Carnegie Learning, Inc.
501 Grant St., Suite 1075
Pittsburgh, PA 15219-4447
sberman@carnegielearning.com

Abstract

We describe preliminary results of the Integrating Human and Automated Tutoring Systems (IHATS) Project, the goal of which is to leverage a unique dataset containing information about student usage of both an automated tutoring system for algebra as well as transcripts of chat sessions between these same students and human tutors. We seek answers to questions about what affective, behavioral, and cognitive factors predict that students will seek human assistance (and/or factors that drive them to human assistance) while using an automated tutoring system and what characterizes when such tutor-tuttee interactions will be enhance learning, using data from the automated tutoring system to measure such learning. The project leverages a variety of statistical and machine learning techniques to answer these questions. Longer term, answers to these questions will be vital to developing systems that can intelligently guide students (i.e., make instructional handoffs) between automated and human sources of assistance as and when necessary.

1 Introduction

Although course designers typically intend to provide a complete set of materials for students to use to meet course objectives, students have always gone beyond these course materials and resources when they perceive that they need assistance beyond that provided in the course. The temptation to go beyond the course materials given is particularly acute in online courses, where, for many topics, there are extensive online text, videos, interactives and assessments available. There have been many approaches to recommending or curating such materials in order to provide students with the best educational experience (Jesukiewicz and Rehak, 2011; Manouselis et al., 2014) and such approaches have had mixed success. However, less attention has been paid to a different kind of choice: when to seek help from another person, particularly an expert teacher or tutor.

Choosing to seek assistance from an expert instructor involves considering different factors from choosing between (non-human) educational resources. Although the effectiveness of human support varies (Cohen, Kulik & Kulik, 1982), the superiority of human tutoring has been well documented (Bloom, 1984; Slavin, 1987; vanLehn, 2011). Humans are better able to diagnose student knowledge gaps and misconceptions, and they are more flexible and interactive in offering explanations and activities that fit with the student's knowledge and preferences (Chi et al., 2001). Humans may also

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

be emotionally supporting in a way that is difficult (or impossible) to achieve in an automated system. For these reasons, duplicating the effectiveness of human tutoring has long been a goal of intelligent tutoring systems (c.f. Corbett, 2001).

However, human tutoring has its downsides. Individual human tutors are very expensive, and even group instruction led by a teacher is often much more expensive than books, videos and other sources of information. In addition, students may hesitate to request help from tutors or instructors because they are embarrassed that they do not understand or fearful that they might ask what is considered a stupid question (Schofield, 1995). They might also be hesitant to utilize human instructional resources because they understand that the instructor's time is limited.

There are many instructional environments in which students need to consider the advantages and disadvantages of asking an expert for help and choose between them. For example, in a blended classroom environment, students working on an intelligent tutoring system in a high school may either ask for help from the tutor or raise their hand and wait for the teacher to answer their questions. In an online environment, students may have the choice to continue working with the course materials, search for additional online materials or consult an online human tutor, including through a chat-based tutoring system like tutor.com.

While intelligent tutoring system developers may have set a goal to meet or exceed the effectiveness of individual human tutors, a better goal might be to consider how to optimize the complete educational environment, including both the tutoring system, other online (and offline) resources and experts (or peers) who may be able to help. Even if we get to the point where intelligent tutoring systems, on their own, are inarguably better than individual human tutors, it is still likely that the combination of both kinds of resources, utilized appropriately, will produce a better educational experience than either one alone. We thus envision a more effective educational system as one where both human and automated instructional resources are available and which includes a recommendation system that can direct students to either use or refrain from using human instructional resources, as appropriate. Such a system would aim to maximize student learning, given the very limited resource represented by the human expert's time.

In order to build such a recommendation system, we need better information about what why students using an automated resource such as an intelligent tutoring system choose to access human resources as well as information about what kinds of interactions with the human tutors are most effective. If we find that students substantially differ in their motives for accessing experts and that we can reliably detect these motives based on their use of automated instructional resources, we will be better able to inform experts about effective instructional or supportive practices. In addition, we should be better positioned to advise students as to whether consulting an expert is advisable at any given time or whether there might be a better and less expensive instructional resource to consult.

A particular interest is the extent to which students are using human experts for affective vs. cognitive reasons. Some preliminary analyses suggest, for example, that a substantial portion of the use of human experts is for affirmation "I think I understand this; am I right?") rather than instruction "Please help me with this topic."). Understanding motives at this level may help us better allocate human experts or find alternatives to using human support.

This paper extends previously-reported data (Ritter et al., 2016) from this dataset.

2 Setting

To better understand why students choose human tutors and how we can make best use of that scarce resource, we have been exploring a unique dataset recording interactions from students in two online developmental mathematics courses offered through an online university. The courses comprise a sequence and cover the topics required for students to progress to college-level mathematics. Each course was five weeks long, with weekly start dates for each course. We sampled a six-month window of students taking the course, resulting in data from 5369 students taking the first course in the sequence, 4659 students taking the second course in the sequence and 6877 students taking both courses in the sequence within our time window. Note that, in the analyses to be reported, we do not aggregate data across courses, so individual students who took both courses would be considered two students in the dataset.

Course requirements included completing work in Carnegie Learning's Cognitive Tutor (CT), a mastery-based intelligent tutoring system. Students were given weekly assignments to complete in CT, and students spent approximately 35 hours using CT during the course. As part of the course, students were also given free and unlimited access to tutor.com (TDC), an online chat-based tutoring service. Of the 16,905 individual students taking the course, 3320 (approximately 16%) chose to use TDC at least once, resulting in 19,248 TDC sessions, with the average session lasting approximately 26 minutes.

Our dataset includes detailed logs of student interactions with CT, including all attempted answers by the students, hint requests and tutor messages and evaluations of correctness, along with timing data. We also have linked chat logs for TDC sessions for all students who accessed that resource.

Since students almost always accessed TDC after substantial time working in CT and also returned to CT after using TDC, we can use the student's behavior in CT prior to using TDC to understand the factors that influence them to seek human tutoring, and we can use CT as a pre- and post-test around the TDC session to understand the impact of the human tutoring session on the student's knowledge and performance.

3 Instructional Technologies

3.1 Carnegie Learning's Cognitive Tutor

Carnegie Learning's Cognitive Tutor (CT; Ritter et al., 2007) was a required component of the two developmental algebra courses we consider in the present work. CT Algebra content was divided into 58 topics for the first algebra course and 44 topics for the second algebra course. Each CT topic is mapped to a set of fine-grained knowledge components (KCs) on which a student demonstrates mastery by solving complex, multi-step problems. In CT's mastery learning approach, once CT estimates that a student has mastered all of the KCs for a topic, the student is moved on to new material in a subsequent topic. Problems within a topic are adaptively selected by the CT to emphasize the set of KCs that a student has yet to master in a topic. Further, context-sensitive hints and immediate feedback are presented to students as they work through the steps of a problem.

Student KC mastery in CT is modeled using the framework of Bayesian Knowledge Tracing (BKT) (Corbett & Anderson 1995). BKT can be viewed as a Hidden Markov Model with two unobserved knowledge states (i.e., each KC either "known" or "unknown") and four parameters for each KC[1]: the probability that a student has prior knowledge of a KC, the probability that a student learns/masters a KC on a particular practice opportunity (i.e., the probability of transitioning from the unknown to the known state), the probability that a student produces a correct response in the unknown state (i.e., guesses correctly), and the probability that a student produces an incorrect response despite being in the known state (i.e., "slipping"). In the Cognitive Tutor's implementation of BKT, the probability of "forgetting" (or transitioning from the known to the unknown state) is set to zero for each KC.

3.2 Tutor.com

Since the two algebra courses under consideration were online courses without an opportunity for face-to-face interaction between students and instructors, the university made one-on-one mathematics tutoring available via the TDC online chat-based tutoring service. Transcripts of tutor-tutee chat sessions that occurred during these math courses were made available to the authors, providing a unique data set that couples human-to-human learning data with human-to-computer learning data (i.e., students interacting with CT). Previous work has analyzed data from these human-to-human tutor-tutee chat transcripts (Samei, et al. 2015), but these datasets did not include information about student learning prior to interacting with a human tutor or subsequent to interacting with human tutors. The availability of this sort of "pre-test" and "post-test" CT data enables us to use statistical and machine learning techniques to answer questions previously unexplored in the ITS literature.

[1]Though outside the scope of this paper, the interested reader is directed to substantial literature that uses a variety of machine learning and statistical methods to personalize BKT parameters and otherwise improve BKT models of student learning using data (e.g., Pardos & Heffernan 2011; Yudelson, et al. 2013; Khajah, et al. 2014). Large amounts of data from the Cognitive Tutor are available via the Pittsburgh Science of Learning Center's DataShop repository (Koedinger, et al. 2011): http://pslcdatashop.web.cmu.edu/.

4 Technical Approach

4.1 Research Questions

Of primary concern to our Advanced Distributed Learning Initiative (ADL)-funded Integrating Human and Automated Tutoring Systems (IHATS) Project are two research questions:

1. What factors predict that a student will seek out human tutoring assistance while they are using an ITS like CT?
2. What characteristics of student chat sessions with human tutors predict that such sessions are educationally effective?

In what follows, we describe machine learning models that allow us to make inferences about CT-related factors that may play important roles in students' decisions to seek out human tutoring as well as natural language processing efforts to infer characteristics of student chat sessions that may produce educationally effective interactions.

4.2 "Detector" Models of Gaming the System, Off-Task Behavior, and Affect

While a variety of features of interest are readily available from CT's logging mechanisms to predict whether students will seek out human tutoring (e.g., % of KCs mastered in a topic, number of hints requested, number of errors made), other behavioral and affective factors can be inferred by machine learning and statistical models that operate on features extracted or "distilled" from these logs. Substantial literature in the educational data mining community and elsewhere is concerned with developing such models. Of particular interest are so-called "detector" models of gaming the system (Baker & de Carvalho 2008; Baker, et al. 2008), off-task behavior (Baker 2007), and student affective states (Baker, et al. 2012) while using CT. Students game the system when they attempt to take advantage of CT affordances (e.g., rapidly seeking hints or guessing) to make progress through material without genuine learning. Affective states for which machine-learned detectors have been developed include frustration, boredom, and confusion, among others.

In general, these detector models produce predictions about whether particular actions (or approximately twenty second clips of student activity) are examples of a particular behavior or a stretch of time in which a student was in a particular affective state. Detectors were originally developed and validated using field observations in real classrooms. However, recent work has also applied these detectors to data from a similar population of students from the course sequence and university and found strong predictors of learning as measured by scores on a final exam (Fancsali 2014, 2015). Moreover, this recent work found evidence that gaming the system behavior, posited as "harmful" to learning in correlational studies (Baker, et al. 2004; Baker, et al. 2008), is indeed a cause of decreased learning in CT. We return to the possibility of inferring causal relationships for the present study later in this work.

4.3 Natural Language Processing: Dialogue Act, Sub-Act, and Mode Classification

To determine the characteristics of effective tutorial dialogues, we map tutorial dialogues into sequences of actions taken by the tutor and the students and then analyze the sessions in terms of action sequences by the tutor and by the student. To this end, we adopt an approach based on the speech-act theory developed in the philosophy of language literature in the 1960s (Austin 1962; Searle 1969). We are concerned to identify dialogue acts, sub-acts, and dialogue modes. While dialogue acts (e.g., "expressive," "assertive," and "prompting" acts) and sub-acts (e.g., "greeting" as an expressive act, "calculation" as an assertive act, and "question" as a prompting act) are associated with particular utterances in a dialogue, dialogue modes refer to general characteristics of series of utterances within tutorial dialogues (e.g., "rapport building," "process negotiation," "problem identification," and "scaffolding"). We seek to answer our second research question about effective tutorial dialogues by considering patterns of dialogue acts and modes and their associations with improved learning when comparing CT work before encountering the human tutor and CT work after a session with a human tutor.

Analysis of the tutorial transcripts began by hand-coding dialogue acts, sub-acts, and modes for 500 randomly selected human tutoring sessions; these hand-coded sessions provide labeled training data

for machine learning classifiers that can then be applied to the entire corpus of tutorial dialogue data. Automated classification of speech/dialogue acts is a well-studied area (Reithinger 1995; Stolcke et al. 2000, etc.) and has also been studied within the specific context of intelligent tutoring systems (Marineau, et al. 2000; Serafin & Di Eugenio 2004). However, identification and automated classification of dialogue modes has received less attention in the literature; one exception is a manual approach to identifying dialogue modes due to Cade, et al. (2008).

The approach to automated dialogue mode classification adopted in the IHATS project is based on the framework of Markov Logic Networks (MLNs) (Domingos & Lowd 2009), a framework that combines probabilistic graphical models with first order logic. The approach is described in detail in Venugopal & Rus (to appear) and depends upon a recently-developed taxonomy of dialogue modes (Morrison, et al. 2014). Our approach differs from traditional "pipeline" approaches to NLP that would start by classifying dialogue acts, then sub-acts, and finally dialogue modes given label acts and sub-acts using classifiers like support vector machine; instead, the MLN-based approach pursues the joint inference task of learning modes, acts, and sub-acts by relying on associations and dependencies among them. Unfortunately, we do not focus on this important and interesting element of the project in the present work.

5 Data Processing

We subset CT logs so that the the samples of TDC and CT users are approximately the same. We also filtered the data to account for students that dropped out early, errors in the product logging and arrived at a dataset with 3121 CT users that used TDC and 3325 that did not.

Our primary interest is effectiveness of human tutoring (TDC) in the context of working within an automated tutor (CT). To address it, we mined the combined CT-TDC logs of 3121 users for patterns of work where student CT work encompasses interacting with TDC tutors. An example of how these patterns were selected is given in Figure 1. We have chunked student activity in CT in terms of work on topics of content within login sessions. Whenever student's CT work overlapped with TDC chat, it was not considered. Only the CT work that encompassed TDC activity without overlapping it in time was considered. In Figure 1, we have highlighted the activity that we considered for the further analyses. Here, a student has three login sessions with the CT. During the first they work on content from topics 1 and 2. In second, they work on topic 2 only. Finally, they work on topics 2 and 3 in the third CT session. Parallel to CT login session they consult TDC during chat session 1. For the patterns of CT work encompassing TDC work we only select whole chunks of work on topic within a CT login session that does not overlap with the TDC chat session. According to this rule and as Figure 1 shows, CT login session 2 was discarded entirely and we only included the last and first topics from CT login sessions 1 and 3 respectively.

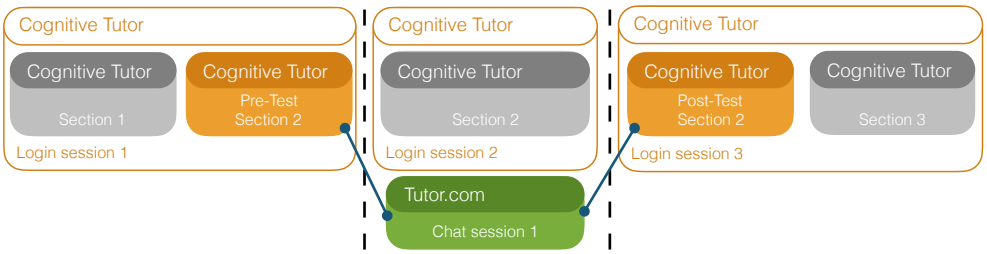

Figure 1: Examples of a pattern of student CT work encompassing interaction with TDC.

We have found 7800 patterns, exemplified in Figure 1, in the data belonging to 2645 out of 3121 students that used TDC. For every pattern, we have summarized student CT work before and after TDC. We selected the sum of student errors and hint requests per problem step as the characteristic of student work and called it assists per step. Additionally, we have recorded the lengths of CT work before and after TDC session, as well as the ratio of topic skills the student mastered by the time that stretch of CT work is over.

We paid special attention to patterns, like the one in Figure 1, where students work on the same topic before and after consulting TDC. We called these cases "overlapping content patterns". There were

2200 out of 7800 patterns of that kind. These cases were special for two reasons. First, since student worked on the same topic, the progress after finishing the TDC session could be directly related to it and measured relative to the work before starting the TDC chat. Second, students that worked on the same material before, potentially during, and after chatting to a human tutor clearly were struggling with that piece of material.

6 Preliminary Results

6.1 Student Retention

We hypothesized that students who were better prepared for the course would be less likely to use TDC than students were were not as well prepared. In order to test this hypothesis, we used student performance on the first module of the course (assigned for completion in the first week) as a measure of course preparation. The hypothesis was supported. Students who used tutor.com had a mean assistance score (total hints plus errors) in the first module of 2824 which opposed to 2028 for students who did not use TDC. The difference was significant (Kruskal-Wallis χ^2=396.6, p-value<0.001).

Nevertheless, students who used TDC at least once were more likely to complete the course (as measured by completion of CT modules) than students who did not use TDC and this effect was consistent across all levels of student preparation (see Figure 2).

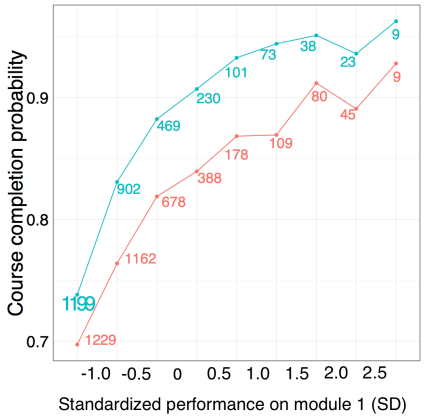

Figure 2: Probability of completing CT work in course, as a function of course preparation.

Over the entire course, students who used TDC did require more assistance to complete material than students who did not use TDC. 41% of TDC users used the service once, and 17% used it twice, with the remaining 42% using it more frequently than this. Our interpretation of these results is that students' lack of course preparation likely encouraged them to use TDC. As a result, they were more likely to remain in the course, even though they continued to struggle more than other students. This interpretation makes sense in light of the fact that most of these students used TDC only once or twice. Such students likely received assistance on only one or two mathematical topics, which is not likely to result in an overall reduction of errors or hint requests within the entire course. The retention effect suggests that TDC may have had a more general affective impact on students. They stayed in the course not because they were better able to correctly complete mathematics topics that were not addressed by TDC but because they were better able to persist in the difficulties they faced in later topics. TDC may have ensured them that someone "had their back" if things got especially difficult.

6.2 Drops in Assistance

To better pinpoint the effect of human tutor help, we have analyzed student work on the granularity level of topics of CT content preceding and following a chat session with a human tutor from TDC. We have estimated the changes in our metric of choice – number of assists (errors and hint requests combined) per problem step and time spent between pre-TDC tutor work in CT and post-TDC student

Table 1: Changes in CT statistics from pre-TDC to post-TDC

Metric	Non-overlapping content 5600 patterns	Overlapping content 2200 patterns
Mean assists/step	3.26 (before) to 3.2 (after)† χ^2=11.6, p-value<0.001***	5.45 (before) to 3.67 (after) χ^2=116.4, p-value<0.001***
Median CT work time, min	31 (before) to 28 (after)	28 (before) to 38 (after)

† We are reporting changes in mean assists per step, however, due to high positive skew, for the statistical test, we have log-transformed the source values to make them more symmetric.

Table 2: Student groups.

			Median duration, min:sec		
Group	Patterns	Students	CT before	TDC	CT after
Confirming. Students that, before consulting TDC mastered $\geq$ 75% of skills in the topic they were working on.	548	425	42:42	22:26	34:06
Learning. Students with <75% of topic skills mastered before consulting TDC and $\geq$10% of skills mastered after that.	609	478	20:23	22:53	47:47
Not learning. Students that mastered 0% of topic skills before consulting TDC and <10% of skills after that.	762	530	24:48	22:58	29:55
Not improving. Students with >0% and <75% of topic skills mastered before consulting TDC and <10% of skills mastered after that.	281	235	42:07	27:25	45:09

work in CT. Special attention was given to the cases when students worked on the same topic before and after using TDC. Results are shown in Table 1.

We have found that, as measured by non-parametric analogue of ANOVA (Kruskal-Wallis rank sum test), the number of assists per step is dropping after consulting the human tutor. In terms of time, pre- and post-TDC CT work lasts around 30 minutes when topic content doesn't overlap. In the case of overlapping content, students spend an average of 38 minutes after ending their human tutoring session.

In order to better understand motivations and effects of use of TDC, we divided students into four groups. "Confirming" students are ones who went to TDC after mastering a substantial number of skills in a topic. We call these students "confirming" because they may be seeking human assistance to confirm their understanding, rather than due to a need to clear up some confusion.[2] "Learning" students are ones who start with low skill mastery and gain mastery subsequent to the TDC session. "Not learning" students are ones who start with no mastery and do not gain much skill mastery in the session following the TDC session. Finally, "not improving" students are ones who start with a moderate level of mastery prior to going to TDC and complete the subsequent session without much improvement. The group definitions and descriptive statistics are given in Table 2.

In Table 2, we can see that, across pattern types, the length of the TDC session fluctuates between 22 and 27 minutes and doesn't change radically. However, the time students invest before and after working with human tutors are quite different. When no learning is detected as per percent skills mastered pre-to-post-TDC time goes from 25 to 30 minutes. Apparently, 30 minutes is not enough to

[2]Future work will test this hypothesis using aforementioned detector models of student affective states, including detectors of confusion.

Table 3: Percent overall TDC time used and course completion for different groups of the 3121 students who used TDC at least once.

Group of TDC users	Overall TDC tutor time used	Avg. course completion (std.err.)†	Minimum time, (hh:mm)	Minimum TDC sessions
Top 1%	15%	79% (4.8%)	25:16	24
Top 5%	40%	75% (2.3%)	9:50	2
Top 10%	55%	73% (1.6%)	5:24	2
Bottom 90%	45%	68% (0.6%)	-	1

† Course completion is defined as topics attempted over all topics available.

transfer whatever help students received into learning achievements in CT. Student learning is most apparent when students spend substantial time in CT subsequent to the TDC session ("learning" students spend almost 48 minutes in the subsequent session). In the cases when students may be using TDC for affirmation, the situation is opposite to the learning case. Namely, the pre-TDC work is longer (43 minutes), while post-TDC work is shorter (34 minutes). The rest of the patterns, in terms of time before and after TDC, gravitate to the learning case with 26 and 44 minutes spent before and after TDC respectively.

6.3 Future Work

So far we have presented preliminary results of a multi-faceted project that leverages several different statistical modeling and machine learning approaches. While we continue to refine analyses to address our research questions, the work presents several questions for (near) future research we now briefly consider.

First, our analysis suggests that some students are "superusers" of relatively costly human tutoring resources. Indeed, the top 1% of TDC users accounted for 15% of all human tutoring time, and the top 10% of human tutoring users accounted for 55% of such time (Table 3) This suggests that modeling the sub-population of superusers has great potential both for better understanding particular student scenarios as they seek human tutoring assistance as well as for cost-savings (without sacrificing learning) if the support that is being sought by the student can be provided by the adaptive tutoring software like CT. Comparing the sub-population of superusers to students who tends to only use the human tutoring resources once or twice can also provide insight to whether differences between these two groups are genuine differences in the kind of support sought (e.g., affirmation and confidence building versus assistance with mathematics content/conceptual difficulties) or just a matter of the degree to which the same kind of support is sought.

The second question helps us address the first question, turning the focus of our modeling efforts from merely predictive aims of determining whether and when a student is likely to seek human help and whether such tutoring will be effective to why a student seeks human help and why such tutoring is effective. These latter questions require causal knowledge that can help us to inform possible interventions that can be implemented within automated tutoring software or as best practices for human tutors during their chat sessions with students. For example, if student affective states like frustration or behavior like gaming the system are predictive of a student's tendency to seek human tutoring help, should we pursue interventions that discourage frustration or gaming the system? If frustration or gaming the system are merely correlated with an underlying cause of students seeking human tutoring assistance (perhaps excessively), then interventions to prevent frustration or gaming to decrease excessive and costly human tutoring assistance could be misguided.

If we have some evidence that these factors are genuine drivers of students use of human assistance, then we can consider interventions to address them. However, the data available to us are only observational, non-experimental data from a real-world implementation of CT with student access to Tutor.com, so we must rely on methods that might help provide causal evidence from such data. The framework of algorithmic search for graphical causal models (Spirtes, et al. 2000) can be brought to bear on this problem and has been used fruitfully on similar datasets of CT use with similar populations of students (Fancsali 2014, 2015). Such models could help to separate features that might be useful in a recommender system for handling instructional handoffs between automated and human tutoring systems from the features which ought to be the target of future interventions. The

future interventions could be designed to decrease particular behaviors and enhance CT in ways that prevent students from ever needing to seek human tutoring help in the first place.

Longer term, we seek to build a recommendation system that will advise students about human and automated forms for assistance, guiding students to the assistance that is most likely to be educationally effective while controlling financial costs. This paper provides preliminary evidence about the factors that are likely to play a part in any such recommendation system; any such system must be based on an understanding of why a student seeks human assistance and what makes relatively costly human assistance educationally effective.

Acknowledgments

This work is supported by a contract from the Advanced Distributed Learning Initiative of the United States Department of Defense (Award W911QY-15-C-0070).

References

Austin, J.L. (1962). How to Do Things with Words. Oxford.

Baker, R.S., Corbett, A.T., Koedinger, K.R., Wagner, A.Z. (2004). Off-task behavior in the Cognitive Tutor classroom: when students "game the system." In Proceedings of ACM CHI 2004: Computer-Human Interaction (pp. 383-390). Vienna, Austria.

Baker, R.S.J.d. (2007). Modeling and understanding students' off-task behavior in intelligent tutoring systems. In Proceedings of ACM CHI 2007: Computer-Human Interaction (pp. 1059-1068). San Jose, CA.

Baker, R.S.J.d., de Carvalho, A. M. J. A. (2008). Labeling student behavior faster and more precisely with text replays. In Proceedings of the 1st International Conference on Educational Data Mining (pp. 38-47). Montreal, Canada.

Baker, R.S., Corbett, A.T., Roll, I., Koedinger, K.R. (2008). Developing a generalizable detector of when students game the system. User Model. User-Adap. 18: 287-314.

Baker, R.S.J.d., Gowda, S.M., Wixon, M., Kalka, J., Wagner, A.Z., Salvi, A., Aleven, V., Kusbit, G.W., Ocumpaugh, J., Rossi, L. (2012). Towards sensor-free affect detection in Cognitive Tutor Algebra. In Proceeding the 5th International Conference on Educational Data Mining (pp. 126-133). Chania, Greece.

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.

Cade, W.L., Copeland, J.L., Person, N.K., D'Mello, S.K. (2008). Dialogue Modes in Expert Tutoring. In Proceedings of ITS 2008, LNCS vol. 5091 (pp. 470–479). Springer.

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

Corbett, A. (2001). Cognitive computer tutors: Solving the two-sigma problem. In User modeling: Proceedings of the Eighth International Conference (pp. 137–147). Berlin, Germany: Springer.

Corbett, A.T., Anderson, J.R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. User Model. User-Adap. 4: 253-278.

Domingos, P., Lowd, D. (2009). Markov Logic: An Interface Layer for Artificial Intelligence. Morgan & Claypool, San Rafael, CA.

Fancsali, S.E. (2014). Causal Discovery with Models: Behavior, Affect, and Learning in Cognitive Tutor Algebra. Proceedings of the 7th International Conference on Educational Data Mining (pp. 28-35). London, England.

Fancsali, S.E. (2015). Confounding Carelessness? Exploring Causal Relationships Between Carelessness, Affect, Behavior, and Learning in Cognitive Tutor Algebra. Proceedings of the 8th International Conference on Educational Data Mining (pp. 508-511). Madrid, Spain.

Jesukiewicz, P. & Rehak, D. (2011). The Learning Registry: Sharing Federal Learning Resources. Presented at the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC).

Khajah, M., Wing, R.M., Lindsey, R.V., Mozer, M.C. (2014). Incorporating latent factors into knowledge tracing to predict individual differences in learning. Proceedings of the 7th International Conference on Educational Data Mining (pp. 99-106). London, England.

Koedinger, K.R., Baker, R.S.J.d., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J. (2011). A Data Repository for the EDM Community: The PSLC DataShop. In: C. Romero, S. Ventura, M. Pechenizkiy, and R.S.J.d. Baker (eds.) Handbook of Educational Data Mining, 43-55. Boca Raton, FL: CRC Press.

Manouselis, N., Drachsler, H., Verbert, K. & Santos, O.C. (2014). Recommender Systems for Technology Enhanced Learning: Research Trends and Applications. Springer Science & Business Media, New York.

Marineau, J., Wiemer-Hastings, P., Harter, D., Olde, B., Chipman, P., Karnavat, A., Pomeroy, V., & Graesser, A. (2000). Classification of speech acts in tutorial dialog. In Proceedings of the workshop on modeling human teaching tactics and strategies at the Intelligent Tutoring Systems 2000 conference (pp. 65–71).

Morrison, D., Nye, B., Samei, B., Datla, V.V., Kelly, C., Rus, V. (2014). Building an intelligent pal from the tutor.com session database phase 1: Data mining. In Proceedings of the 7th International Conference on Educational Data Mining.

Pardos, Z.A., Heffernan, N.T. (2011). Kt-idem: Introducing item difficulty to the knowledge tracing model. In Proceedings of User Modeling, Adaption, and Personalization (UMAP 2011), LNCS vol. 6787 (pp. 243–254). Springer.

Reithinger, N. (1995). Some experiments in speech act prediction. In Proceedings of the AAAI Spring Symposium on Empirical Methods in Discourse.

Ritter, S., Anderson, J.R., Koedinger, K.R., & Corbett, A. (2007) The Cognitive Tutor: Applied research in mathematics education. Psychonomics Bulletin & Review, 14(2), pp. 249-255.

Ritter, S., Yudelson, M., Fancsali, S., Berman, S. (2016) Towards Integrating Human and Automated Tutoring Systems. In Proceedings of the 9th International Conference on Educational Data Mining (EDM 2016), Raleigh, NC.

Samei, B., Rus, V., Nye, B., Morrison, D.M. (2015). Hierarchical Dialogue Act Classification in Online Tutoring Sessions. Proceedings of the 8th International Conference on Educational Data Mining (pp. 600-601). Madrid, Spain.

Schofield, J.W. (1995). Computers and Classroom Culture. Cambridge University Press, Cambridge, UK

Searle, J.R. (1969). Speech Acts: An Essay in the Philosophy of Language. Cambridge University Press.

Serafin, R., Di Eugenio, B. (2004). Flsa: Extending latent semantic analysis with features for dialogue act classification. In Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume (pp. 692–699). Slavin, R. E. (1987). Making Chapter 1 make a difference. Phi Delta Kappan, 69(2): 110–119.

Spirtes, P., Glymour, C., Scheines, R. (2000). Causation, Prediction, and Search. 2nd edition. Cambrige, MA: MIT Press.

Stolcke, A., Coccaro, N., Bates, R., Taylor, P., Van Ess-Dykema, C., Ries, K., Shriberg, E., Jurafsky, D., Martin, R., Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. Comput. Linguist. 26(3): 339–373.

vanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist 6(4): 197-221.

Venugopal, D., Rus, V. (to appear). Joint Inference for Mode Identification in Intelligent Tutoring Systems. Proceedings of the 26th International Conference on Computational Linguistics. Osaka, Japan.

Yudelson, M., Koedinger, K., Gordon, G. (2013) Individualized Bayesian Knowledge Tracing Models. In: Lane, H.C., and Yacef, K., Mostow, J., Pavlik, P.I. (eds.) Proceedings of 16th International Conference on Artificial Intelligence in Education (AIED 2013), Memphis, TN. LNCS vol. 7926, (pp. 171-180). Springer-Verlag Berlin Heidelberg.