Beyond the *code-and-count* analysis of tutoring dialogues [1]

Stellan OHLSSON [a], Barbara DI EUGENIO [a,2], Bettina CHOW [a], Davide FOSSATI [a], Xin LU [a], and Trina C. KERSHAW [b]

[a] *University of Illinois at Chicago, IL, USA*

[b] *University of Massachusetts at Dartmouth, MA, USA*

Abstract. In this paper, we raise a methodological issue concerning the empirical analysis of tutoring dialogues: The frequencies of tutoring moves do not necessarily reveal their causal efficacy. We propose to develop coding schemes that are better informed by theories of learning; stop equating higher frequencies of tutoring moves with effectiveness; and replace ANOVAs and chi-squares with multiple regression. As motivation for our proposal, we will present an initial analysis of tutoring dialogues, in the domain of introductory Computer Science.

Keywords. Tutoring dialogues, annotation, multiple regression

1. Introduction

In this paper, we raise a methodological issue concerning the empirical analysis of tutoring dialogues, in general and in the service of building dialogue interfaces to Intelligent Tutoring Systems (ITSs). Like many others [2,7,8,10,13,19], we have been engaged in the collection and analysis of one-on-one tutoring dialogues [4,5]. The goal is to glean insight into why tutoring is effective, and to build computational models of effective tutoring strategies, which in turn are an essential component of dialogue interfaces to ITSs. However, even with so much effort, the community does not yet agree on a repertoire of effective tutoring strategies. We believe this is because everybody, including ourselves, has been equating effectiveness with frequency: namely, what tutors, in particular expert tutors, do most often is what is deemed to be effective. However, frequency per se does not prove effectiveness, as we will explicate below. In addition, coding categories for tutoring dialogues have often been developed bottom-up, certainly informed by linguistics theories of speech acts, but not directly informed by theories of learning. In this paper, we present our argument, and we follow up with two proposals: develop coding schemes that are better informed by theories of learning, and replace what we call the *code-and-count* methodology with an analysis based on multiple regression. As motivation for our proposal, we will present an initial analysis of tutoring dialogues, in the domain of introductory data structure and algorithms in Computer Science.

[1] This work is funded by awards N000140010640 and N000140710040 from the Office of Naval Research, and in part by awards IIS–0133123 and IIS–0513553 from the National Science Foundation.

[2] Corresponding Author: Computer Science (M/C 152), 851 S. Morgan St., Chicago, IL 60607, USA.

2. The analysis of tutoring dialogues: A critique

2.1. Coding Categories

Tutor-tutee interactions are so rich and complex that researchers have not yet converged on a shared set of tutoring moves, i.e., the basic units of analysis. For example, Person [16] lists 23 categories of tutoring behaviors, including *hint, prompt, pump, bridge, summarize, ask clarification questions, ask comprehension-gauging questions, provide counterexamples, give direct instruction, force a choice, provide a preview, provide examples, complain, revoice*. Although long, this list is not inclusive. For example, VanLehn et al. [19] coded tutoring transcripts for the occurrence of *impasses*, *explanations* and *hints about subgoals*; the first and third of these are not identical to any category in Person's list. The coding categories from [7] overlap with those mentioned so far, e.g. hints, and so do our own, e.g., prompting, [5], but do not coincide. The point is that tutoring research has not converged on a widely agreed-upon theory of how the behavior of tutors should be segmented and characterized, a necessary preliminary step to investigating which parts of tutoring produce the high learning gains. The various lists of tutoring moves have sometimes been created in a bottom-up fashion, in response to the contents of the transcripts collected by a research team. Another influence is to be found in linguistic theory, especially theories of dialogue acts, and yet another in prior pedagogical concepts about how tutoring might work (e.g., Socratic questioning). To date, attempts to answer the question of why tutoring is effective have been less responsive to what we know about how people learn. Yet, instruction can only work by supporting learning, so considering how people learn is a natural starting point [14]; a core coding scheme of this sort can be later enriched with other categories, i.e. those suggested by theories of dialogue acts, if necessary.

2.2. The code-and-count *methodology*

The question of how tutoring strategies and moves should be conceptualized interacts with the standard methodology used in empirical tutoring studies. The latter often proceed by classifying tutoring behaviors and then assessing which of those behaviors are the most important in achieving the superior learning outcomes associated with tutoring by counting the frequency of each behavior in a corpus of transcripts. This code-and-count methodology can only be as informative as the coding categories; if the latter do not categorize tutoring moves in a way that corresponds to the actual workings of tutoring, then the frequencies of those categories will be uninformative.

Even if a code-and-count study began with a principled set of categories, there are other methodological problems that blunt the impact of such studies. It rests on the assumption that the tutoring behaviors that account for the superior learning gains obtained with tutoring would turn out to be those behaviors in which the tutors engage frequently. Hence, code-and-count would generate the right tutoring theory by revealing the highest-frequency tutoring moves. In retrospect, this assumption (made by everybody in the community, including ourselves) appears to us to be flawed. There is no guarantee that the moves that account for most of the variance in learning outcomes are necessarily the moves that occur most frequently. After all, the tutors themselves, even expert tutors, do not have a theory of tutoring (or else we could accomplish our objective by asking them)

– even if some of them, for example CIRCSIM experienced physiology tutors, constantly reflect on and evaluate their own tutoring [12]. Furthermore, human interactions are very complicated and shaped by multiple factors, including the standard interaction patterns of the surrounding culture and the degree and nature of the rapport between a particular tutor and a particular tutee. Hence, the maximally effective tutoring moves – the ones most worthwhile to mimic in an artificial tutoring system – might be few and embedded within a lengthy interaction that is rich, varied and complicated for a variety of reasons that bear little causal relation to the learning outcome. Compare the situation to negotiation dialogues: If two negotiators after a lengthy discussion reach a particular agreement, which utterances in their transcript were causally related to the outcome? Clearly, it is not necessarily the type of utterance that was most frequent. A more sophisticated approach is to code-and-count tutoring dialogues produced by expert tutors, specifically. It is reasonable to assume that expert tutors will do more of whatever it is that works in tutoring, so the most frequent types of tutoring moves by expert tutors should have a high probability of cutting close to the tutorial bone. However, even a study of expert tutors can only be as revealing as the set of categories is insightful, and there is no guarantee that the highest frequency moves are those with the strongest impact on learning.

In addition, studies of expert tutors have weaknesses of their own. As Person recently documented [16], only a few expert tutors have been studied [3,7,11], in part because the total set of studies of expert tutors is small and in part due to the fact that the very same expert tutors have appeared in more than one study. In addition, tutors are often identified as 'expert' on the basis of indirect indicators such as how long they have been tutoring, how much they are paid, etc. With respect to the goal of identifying the dimensions of tutoring that are causally related to high learning gains, the operative concept is not *expert* but *effective* tutors, and the measure of effectiveness is some measure of learning outcomes, not any property of the tutors themselves. If some of the expert tutors that have been studied do not in fact achieve high learning gains, their presence in the data base constitutes noise. In the beginning of our research on expert vs. non expert tutors [5,6] we thought to overcome such weaknesses by (a) verifying that our expert tutors did indeed produce higher learning gains by using pre- and posttests, and (b) by looking at *differences* between more and less effective tutors. In other words, we thought that the differential frequencies that would turn up in the code-and-count would indicate whatever it is that effective tutors do more often than ineffective tutors, and those types of behaviors would be good candidates for the tutoring moves that produce the better learning outcomes. This same approach was taken by the CIRCSIM-Tutor group [7,9]. Although we believe that these two methodological improvements – measure learning gains and look at differential rather than absolute frequencies – are important, we now recognize that they do not completely overcome the weaknesses that we have identified above: the creation of a coding system that does not take into account theories of learning and the assumption that the effective tutoring moves are necessarily more frequent than those with less causal impact on learning. In addition, as we will discuss below, a tutor who has been deemed *expert* according to some a priori criteria may not be more effective than a tutor who has not been deemed as expert.

3. Proposed Solution

3.1. Coding Categories

The system for categorizing tutoring behavior should be informed by what we know about learning (as well as by the manifest content of the relevant tutoring transcripts). Although no theory of skill acquisition is widely accepted, most researchers would agree that people learn in at least the following four ways:

1. People can learn by capturing and encoding successful steps during problem space search.
2. People can learn by detecting and correcting their errors.
3. People can learn by encoding declarative facts about the domain or about the types of problems or tasks they are practicing, delivered via discourse.
4. People can learn by compiling declarative representations of the tactics and strategies they are trying to learn into executable cognitive strategies.

The claim is not that this is the complete theory for how people learn, only that people learn in at least these four ways. Each of these types of learning suggests a particular way in which instruction can support learning:

1. A tutor can provide positive feedback to confirm that a correct but tentative student step is in fact correct.
2. A tutor can provide negative feedback that helps a student detect and correct an error.
3. A tutor can state the declarative information about the domain.
4. A tutor can tell the student how to perform the task.

For each of these types of tutorial inputs, we can explain why and how they might support learning in terms of current cognitive theories of skill acquisition [1,15,17,18]; and they have all been documented as occurring in tutoring dialogues. It seems reasonable, then, to code tutoring transcripts for the occurrence of (at least) these four categories of tutoring behavior. Note that the empirical methodology that we propose below will signal to us whether these categories are sufficient or we need to add other categories.

3.2. Multiple regression

We propose to move away from absolute frequencies to a different criterion for how to decide whether a category of tutoring behavior is causally related to learning: Multiple regression of learning outcomes per tutoring session onto the frequencies of the different tutoring moves. First, we intend to shift the unit of analysis from *tutor* to *tutoring session*. Past research implicitly assumed that tutoring expertise is general across recipients; that is, if a tutor is effective with student X, he/she tends to be effective with student Y as well. Therefore, we can uncover the moves of effective tutoring by statistical aggregation of code-and-count data over students and sessions but within tutor. However, the richness of human dialogues once again intervenes. It is highly likely that each tutor succeeds better with some students than with others; better rapport, more closely related and matching linguistic habits or thoughts, and so forth. It is therefore reasonable to focus on tutoring sessions, not tutoring persons. The question then becomes, what happens in those sessions in which much learning happened, and what differentiates them from those session in which it didn't? This is clearly a different question than asking what certain persons, expert tutors, tend to do that other persons do not do, or do less of. The next methodological innovation, which develops a trend that is already present in some recent tutoring studies [19], is to move away from ANOVAs and chi-squares to a correlational approach. We anticipate to see a wide variety of learning outcomes across sessions, and there may be wide variations in the frequencies of the four theory-based tutoring categories outlined above. To answer the question which of these categories are causally related to the learning outcomes, we will employ multiple regression, with the amount of learning per session as the predicted variable and the frequencies of the tutoring behaviors as the predictor variables. The beta weights, the partial regression coefficients, will provide information regarding the relative strength of the relations between the relevant tutoring moves and the learning outcomes. This method still relies on the frequency of each type of tutoring move as an important variable, but it does not assume that the more effective tutoring moves are necessarily more frequent, in absolute numbers, than the less effective ones. The method reveals whether *variation* in the frequency of one type of tutoring move is more strongly related to *variation* in learning outcomes than another type of move, regardless of which type of move is more or less frequent. An additional advantage of the correlational methodology is that it will tell us if we are on the wrong track: One possible outcome is that none of the partial regression coefficients are significant. This is a sign that the categories of tutoring moves are not the right ones, and that the transcripts need to be re-coded with a different set of categories (i.e., that the theory behind the category system is wrong and needs to be replaced by a different learning theory). There is no counterpart to this in the code-and-count methodology: Any set of codes will always generate some frequencies, and there will always be some code that turns out to be more frequent than another. There is no built-in warning signal that the category system is fundamentally flawed, but in the correlational approach, low and non-significant correlations do provide such a warning.

4. An initial analysis of tutoring dialogues in Computer Science

Our current goal is to investigate computational models of tutoring dialogues in order to build a dialogue interface to an ITS for basic data structures and algorithms. We have thus engaged in an extensive collection of Computer Science tutoring dialogues. At the moment, we have data from 82 subjects altogether, 28 control and 54 tutored. At the time of writing, we have transcribed 80% of the dialogues and started to code those already transcribed. Hence, we are not in a position to report a complete multiple regression analysis. However, we include here some preliminary findings that show the need for this kind of analysis. The subjects were recruited from CS introductory courses. They were either (a) tutored by a very experienced tutor (CS professor with 30 years experience in small liberal art schools), (b) tutored by a less experienced tutor (a CS senior with just few hours of "helping friends" under his belt), or (c) not tutored (control condition). They were given a 15 minute pretest immediately before the tutoring session, and the same posttest immediately afterward. The test contains eight problems, divided into three groups: problems 1 and 2 on linked lists, problems 3 and 4 on stacks, and problems 5-8 on binary search trees (BSTs). The students in the control condition took the same tests but in a classroom setting; instead of being tutored, they read appropriate excerpts of the textbook, at their own pace. The tutoring sessions lasted at most 40 minutes, with some advanced students finishing a bit earlier. The pre- and posttests were graded by two graders who were blind to condition. Each test problem was worth 5 points, for a total of 40 points. Grader differences of 2 points or more were resolved by discussion, smaller grader differences by averaging over the two graders. The measure of learning gain was the difference between the post- and pretest scores. Initial general findings on learning within groups are that (these results are based on paired sample t-tests, where the two variables of interest are the scores on the pre- and posttest):

- students learn significantly in all conditions. Specifically, both CS majors and non CS majors learn significantly in each problem
- students in the control condition learn significantly on half of the problems, i.e. problems 3, 4 (stacks), and 5, 6 (BSTs)
- students with the less experienced tutor learn significantly on 5 problems out of 8, i.e. problems 3, 4 (stacks), and 5, 6, 7 (BSTs)
- students with the more experienced tutor learn significantly on each problem

Further, comparing students across different conditions, we find that (all these results are based on ANOVAs, followed by Bonferroni tests):

- both tutors are effective, i.e., subjects tutored by either tutor learn more than subjects in the control condition
- even if subjects learn a bit more with the more experienced tutor, there is no significant difference between the subjects tutored by one tutor or the other
- subjects who are not CS majors learn more than subjects who are CS majors
- there are differences for some selected problem regarding learning gain between tutored and control subjects, and between CS and non-CS majors

To start with our proposed methodology, we ran three multiple regressions, one for each subject matter (linked lists, stacks, binary search trees). Of concern are three potential predictor variables that are of interest in understanding what happens during tutoring: the students prior knowledge (pretest score), the level of experience of the tutor (less, more), and the time spent during tutoring on problems pertaining to each subject matter topic (time on task). We used both the posttest score and the gain score as measures of learning. We report the posttest score analysis. (The gain score analysis gave the same results.) We report the results of the multiple regression in Table 1. We found that the experience of the tutor did not significantly affect the posttest scores. More specifically, for linked lists, higher pretest scores and increased time on task predicted higher posttest scores: pretest score accounts for 31% of the variance, and time on task for 6%; for stacks, higher pretest scores predicted higher posttest scores, accounting for 50% of the variance; for trees, higher pretest scores predicted higher posttest scores, accounting for 18% of the variance.

These preliminary results suggest the following. First, we notice that the experience level of the tutor did not affect average learning gains, nor did tutor experience explain a significant portion of the variance in the learning gains on any of the three subject matter topics. This reinforces our contention that we should analyze tutoring dialogues by focusing not on the participant (i.e., tutor, or type of tutor or student), but rather, on the session as unit of analysis. Whereas the distinction between more and less expert tutors, which we subscribed to in our previous work, is an appealing one, it does not

		R^2	β	p
Linked lists	Pretest	0.308	0.635	< 0.001
	Tutor experience	0.001	0.033	ns
	Time on task	0.063	0.294	0.009
Stacks	Pretest	0.505	0.665	< 0.001
	Tutor experience	0.002	0.041	ns
	Time on task	0.006	-0.061	ns
BSTs	Pretest	0.180	0.483	< 0.001
	Tutor experience	0.078	-0.115	ns
	Time on task	0.025	0.234	ns

Table 1. Results from multiple regression, by topic

take into account the other half of the equation, i.e., the student. This observation does not exclude that a more experienced tutor is likely to have more successful sessions, and hence, be more effective on the whole, than a less experienced tutor [5,9]. Second, the strongest predictor of posttest score is the pretest score. Time on task comes in as second best. These are not surprising, but common findings in educational research. The point for present purposes is: The variation in the learning gains is due to many factors, each factor being responsible, so to speak, for a portion of that variance. The more variance is explained by one variable, the less is left to be explained by another. Any claim to the effect that such and such a tutoring move is causally related to learning gains must show that variations in the use of that tutoring move explains variance in learning gains, over and above what is explained by other variables such as prior knowledge and time on task. Unless the methodology for analyzing tutoring dialogues takes such variables into account, the empirical support in favor of a hypothesized tutoring move is weakened. Third, even after we take these variables into account, depending on topic there is still between 80% and 50% of the variance in the posttest scores left to explain. The obvious hypothesis is that this portion of the variance is indeed due to variations in how the tutor-tutee interaction went in any one tutoring session. To assess this hypothesis, the analysis needs to be repeated with the frequencies of the tutoring moves of interest as predictor variables.

5. Conclusions

To analyze tutoring dialogues with the purpose of extracting a theory of tutoring that can inform the design of ITSs is to ask which of the many and varied actions that tutors take in the course of interacting with a student is causally related to the high learning gains that attract us to the tutoring scenario in the first place. Frequency does not prove causal efficacy. We believe this is why code-and-count studies based on bottom-up coding schemes have so far failed to resolve the issue of what makes tutoring effective. We are developing an alternative methodology, and the purpose of this paper is to put that methodology before the ITS research community for comment and criticism. The methodology has the following parts: First, coding schemes should start from what is known about learning. A tutoring move should be identified in terms of which type of learning they support. Second, data analysis should use the individual tutoring session as the unit of analysis, not the tutor. Third, the analyses of tutoring dialogues should include data from all the variables that are known to affect educational outcomes, such as prior knowledge, time on task, etc. Fourth, the variable of interest is not any property of the tutors per se such as number of years of tutoring experience, but the learning outcomes. Fifth, to prove that a tutoring move is effective is to prove that it accounts for variance in learning gains over and above the variance explained by such base line variables. One tool to accomplish this is multiple regression.

References

[1] John R. Anderson. Knowledge compilation: The general learning mechanism. In R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, editors, *Machine Learning*, volume 5, pages 289–310. Los Altos, CA: Kaufmann, 1986.

[2] Michelene T. H. Chi, Stephanie A. Siler, Takashi Yamauchi, and Robert G. Hausmann. Learning from human tutoring. *Cognitive Science*, 25:471–533, 2001.

[3] Teresa del Soldato and Benedict du Boulay. Implementation of motivational tactics in tutoring systems. *Journal of Artificial Intelligence in Education*, 6(4):337–378, 1995.

[4] Barbara Di Eugenio, Davide Fossati, Dan Yu, Susan Haller, and Michael Glass. Natural Language Generation for Intelligent Tutoring Systems: a case study. In *AIED 2005, the 12th International Conference on Artificial Intelligence in Education*, Amsterdam, The Netherlands, 2005.

[5] Barbara Di Eugenio, Trina C. Kershaw, Xin Lu, Andrew Corrigan-Halpern, and Stellan Ohlsson. Toward a computational model of expert tutoring: a first report. In *FLAIRS06, the 19th International Florida AI Research Symposium*, Melbourne Beach, FL, 2006.

[6] Barbara Di Eugenio, Xin Lu, Trina C. Kershaw, Andrew Corrigan-Halpern, and Stellan Ohlsson. Positive and negative verbal feedback for intelligent tutoring systems. In *AIED 2005, the 12th International Conference on Artificial Intelligence in Education*, Amsterdam, The Netherlands, 2005. (Poster).

[7] Martha Evens and Joel Michael. *One-on-one Tutoring by Humans and Machines*. Mahwah, NJ: Lawrence Erlbaum Associates, 2006.

[8] Barbara A. Fox. *The Human Tutorial Dialogue Project: Issues in the design of instructional systems*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1993.

[9] Michael Glass, Jung Hee Kim, Martha W. Evens, Joel A. Michael, and Allen A. Rovick. Novice vs. expert tutors: A comparison of style. In *MAICS-99, Proceedings of the Tenth Midwest AI and Cognitive Science Conference*, pages 43–49, Bloomington, IN, 1999.

[10] Arthur C. Graesser, Natalie K. Person, and Joseph P. Magliano. Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology*, 9:495–522, 1995.

[11] M. R. Lepper, M. F. Drake, and T. O'Donnell-Johnson. Scaffolding techniques of expert human tutors. In K. Hogan and M. Pressley, editors, *Scaffolding student learning: Instructional approaches and issues*. Cambridge, MA: Brookline, 1997.

[12] Joel A. Michael and Allen A. Rovick. *Problem-solving in physiology*. Prentice Hall, Upper Saddle River, NJ, 1999.

[13] Johanna D. Moore, Benoît Lemaire, and James A. Rosenbloom. Discourse generation for instructional applications: Identifying and exploiting relevant prior explanations. *Journal of the Learning Sciences*, 5(1):49–94, 1996.

[14] Stellan Ohlsson. System hacking meets learning theory: Reflections on the goals and standards of research in artificial intelligence and education. *International Journal of Artificial Intelligence and Education*, 2(3):5–18, 1991.

[15] Stellan Ohlsson. Learning by specialization and order effects in the acquisition of cognitive skills. In E. Ritter, J. Nerb, T. O'Shea, and E. Lehtinen, editors, *In order to learn: How the sequence of topics affects learning*. New York, NY: Oxford University Press, 2007. To appear.

[16] Natalie Person. Why study expert tutors? Presentation at ONR Contractors' Conference on Instructional Strategies, February 2006.

[17] R. Sun, P. Slusarz, and C. Terry. The interaction of the explicit and the implicit in skill learning: A dual-process approach. *Psychological Review*, 112:159–192, 2005.

[18] Kurt VanLehn. Learning by explaining examples to oneself: A computational model. In S. Chipman and A. L. Meyrowitz, editors, *Foundations of knowledge acquisition*, pages 25–82. Boston: Kluwer, 1993.

[19] Kurt VanLehn, Stephanie Siler, and Chaz Murray. Why do only some events cause learning during human tutoring? *Cognition and Instruction*, 21(3):209–249, 2003.