Shahriar, T., & Matsuda, N. (Accepted) What and how you explain matters: Inquisitive Teachable Agent Scaffolds Knowledge-building for Tutor Learning. In N. Wang, G. Rebolledo-Mendez, O. C. Santos, V. Dimitrova & N. Matsuda (Eds.), Proceedings of the International Conference on Artificial Intelligence in Education: Springer.

What and how you explain matters: Inquisitive Teachable Agent Scaffolds Knowledge-building for Tutor Learning

Tasmia Shahriar [0000-0003-0199-7757] and Noboru Matsuda [0000-0003-2344-1485]
North Carolina State University, Raleigh NC 27695, USA
(tshahri,noboru.matsuda)@ncsu.edu

Abstract. Students learn by teaching a teachable agent, a phenomenon called tutor learning. Literature suggests that tutor learning happens when students (who tutor the teachable agent) actively reflect on their knowledge when responding to the teachable agent's inquiries (aka knowledge-building). However, most students often lean towards delivering what they already know instead of reflecting on their knowledge (aka knowledge-telling). The knowledge-telling behavior weakens the effect of tutor learning. We hypothesize that the teachable agent can help students commit to knowledge-building by being inquisitive and asking follow-up inquiries when students engage in knowledge-telling. Despite the known benefits of knowledge-building, no prior work has operationalized the identification of knowledge-building and knowledge-telling features from students' responses to teachable agent's inquiries and governed them toward knowledge-building. We propose a Constructive Tutee Inquiry that aims to provide follow-up inquiries to guide students toward knowledge-building when they provide a knowledge-telling response. Results from an evaluation study show that students who were treated by Constructive Tutee Inquiry not only outperformed those who were not treated but also learned to engage in knowledge-building without the aid of follow-up inquiries over time.

Keywords: Learning by teaching, teachable agents, tutor-tutee dialogue, knowledge-building, algebra equation solving

1 Introduction

A teachable agent (TA) is a computer agent that students can interactively teach. Literature affirms that students learn by teaching the TA, a phenomenon called the *tutor learning* [1-3]. In this paper, we address students who teach a TA as *tutors*. The current literature suggests that tutors learn by teaching when they reflect on their understanding [4], revisit the concepts, provide explanations to make sense of solution steps [5], and recover from misconceptions or knowledge gaps. These activities are defined as *knowledge-building* activities and are known to facilitate tutor learning better than *knowledge-telling* activities [6, 7], which means rephrasing known information and explanations with shallow reasoning.

Despite the effectiveness of knowledge-building, tutors are often biased towards delivering what they know to their TA (i.e., knowledge-telling) [7-9]. Such a lack of cognitive effort weakens tutor learning. Roscoe *et al.* [1] further argued that tutors only

engage in knowledge-building when they realize they do not know or understand something. Researchers reported that tutors are likely to realize their knowledge gaps and engage in knowledge building if the TA asks follow-up inquiries [6, 7, 10].

In this work, we implement a method for producing TA follow-up inquiries that is capable of engaging tutors in knowledge-building. We call it *Constructive Tutee Inquiry* (CTI). We investigate features of tutors' responses to identify if the tutor engaged in knowledge-building or knowledge-telling. This is the first work to operationalize the detection of knowledge-building and knowledge-telling from tutors' responses and generate follow-up inquiries to guide tutors toward knowledge-building when they fail to do so. To evaluate the effectiveness of the proposed CTI, we address the following research questions: **RQ1**: Can tutor responses be accurately classified into knowledge-building and knowledge-telling to drive CTI? **RQ2:** If so, does CTI facilitate tutor learning? **RQ3:** Does CTI help tutors learn to engage in knowledge-building? An empirical evaluation study was conducted as a randomized controlled trial with two conditions to answer those research questions. In the treatment condition, the TA asked follow-up inquiries (i.e., CTI), whereas, in the control condition, the TA did not ask follow-up inquiries. The results show that treatment tutors outperformed control tutors and learned to engage in knowledge-building over time.

Our major contributions are: (1) we operationalized knowledge-building and knowledge-telling and developed a machine learning model to automatically identify knowledge-telling responses, (2) we developed CTI that guides tutors to commit knowledge-building responses, (3) we conducted an empirical study to validate CTI and demonstrated its effect on engaging tutors in knowledge-building. (4) we built and open-sourced the response classifier and the coded middle-grade tutors' response data.

2 SimStudent: The Teachable Agent

SimStudent [11] is our teachable agent (TA). It is shown at the bottom left corner of **Fig 1** which displays the entire user interface of the online learning environment called APLUS (Artificial Peer Learning Using SimStudent). In the current work, SimStudent is taught how to solve linear algebraic equations. A tutor may enter any equation in the tutoring interface (**Fig 1-a**). Once the equation is entered, SimStudent consults its existing knowledge base and performs one step at a time. A step consists of either choosing one of four basic math transformations to apply (add, subtract, divide, and multiply) or executing a suggested transformation (e.g., actually adding two terms). The knowledge base consists of production rules automatically composed by SimStudent using inductive logic programming to generalize solutions demonstrated by the tutor [12]. When SimStudent suggests a step, the tutor must provide corrective feedback (yes/no) based on their opinion. When SimStudent gets stuck, the tutor needs to demonstrate the step. Over the course of a tutoring session, SimStudent continuously modifies the production rules in its knowledge base according to the tutor's feedback and demonstrations.

The tutor interacts with SimStudent using the chat panel (**Fig 1-b**). Currently, SimStudent asks: (1) "Why did you perform *[transformation]* here?" when the tutor demonstrates a *transformation*, and (2) "I thought *[transformation]* applies here. Why do you think I am wrong?" when the tutor provides negative feedback to SimStudent's

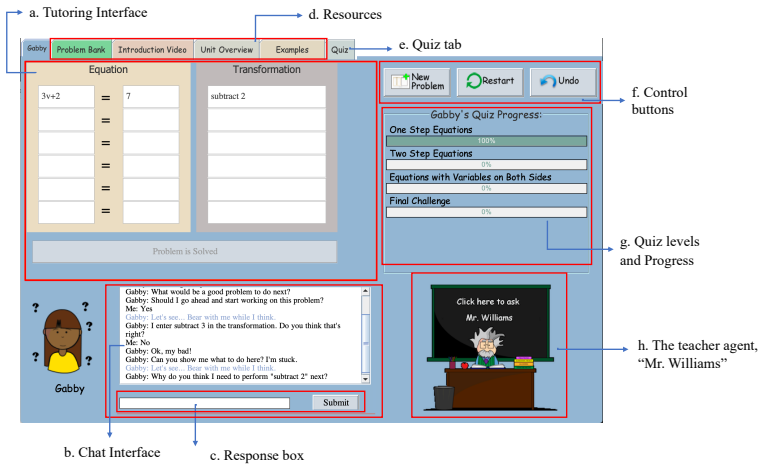

Fig 1: APLUS interface with SimStudent named Gabby in the bottom left corner.

suggested *transformation*. The tutor provides a textual explanation in the response box (**Fig 1-c**) to answer SimStudent's inquiries or may submit an empty response. In any case, SimStudent replies, "Okay!" and proceeds to the next solution step.

The tutor can give a new equation using the "New Problem" button, instruct the TA to solve a problem from the beginning using the "Restart" button, and undo using the "Undo" button (**Fig 1-f**). Additionally, the tutor may quiz (**Fig 1-e**) SimStudent anytime to check its knowledge status. Quiz topics include four levels (**Fig 1-g**). The final challenge consists of equations with variables on both sides. The tutor can refresh their knowledge on solving equations using the resources like the Problem Bank, the Introduction Video, the Unit Overview, and worked-out Examples (**Fig 1-d**). The teacher agent, Mr. Williams (**Fig 1-h**), provides on-demand, voluntary hints on how to teach. For example, if the tutor repeatedly teaches easy one-step equations that he knew prior to using APLUS, Mr. Williams might provide the hint, "Your student failed on the two-step equation. Teaching similar equations will help him pass that quiz item".

3 Constructive Tutee Inquiry

3.1 Motivation

The purpose of Constructive Tutee Inquiry is to help tutors generate knowledge-building responses to TA's inquiries. Existing literature argues that tutors switch from providing knowledge-telling to knowledge-building responses by hitting impasses or moments when they realize they do not know something or need to double-check their understanding [7, 9, 13]. Such moments are highly improbable to attain without proper scaffolding on the tutors' responses.

In the rest of the paper, we call the TA's first inquiry in an attempt to understand a topic the *initial inquiry*. For instance: "Why did you perform *[transformation]* here?" is an initial inquiry where SimStudent attempts to understand the reason behind applying the particular *transformation*. We call the TA's subsequent inquiries after the initial inquiry on the same topic the *follow-up inquiries*.

Constructive Tutee Inquiry (CTI) is a TA's follow-up inquiry to solicit knowledge-building responses. The conversational flow chart using the CTI engine is shown in **Fig 2**. Two major technologies (together named as CTI Engine) that drive CTI are the *response classifier* and the *dialog manager*. The response classifier analyzes a tutor's response and classifies it as one of the pre-defined types that show what information the TA needs to seek in the next follow-up inquiry. The dialog manager selects the appropriate follow-up inquiry based on the tutor's response type with the intention to solicit a knowledge-building response. **Fig 3** illustrates an example inquiry generated by the dialog manager based on the response classifier output class.

3.2 Response classifier

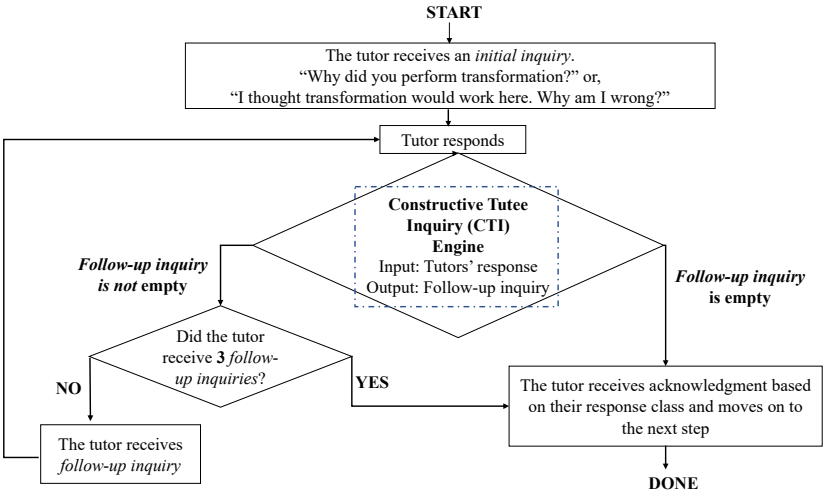

Fig 2: A complete conversation flow chart using CTI Engine

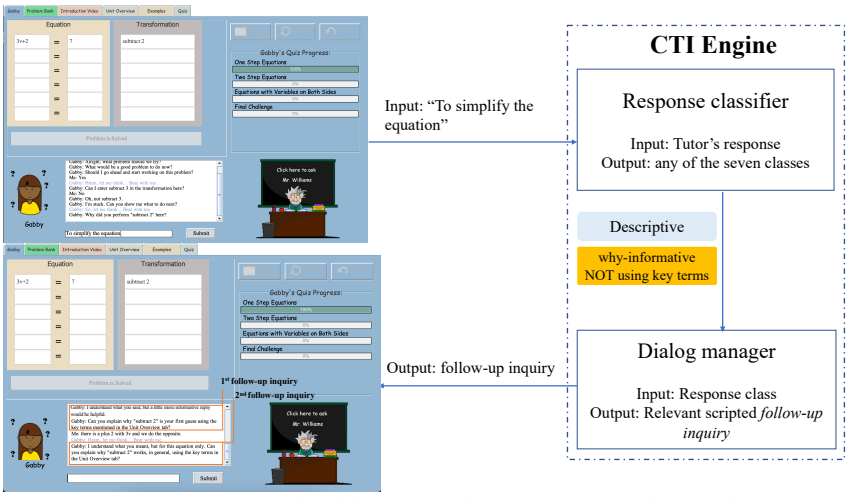

Fig 3: How CTI Engine works with an example inquiry generated by Dialog Manager

The goal of a response classifier is to identify knowledge-building and knowledge-telling responses and to determine what information to seek in the subsequent follow-up inquiry to solicit a knowledge-building response. The theoretical definition of knowledge-building and knowledge-telling responses (described earlier) inspired us to categorize responses in a top-down approach based on *sentence formation*, *relevancy*, *information content,* and *intonation*.

(i) *Sentence formation (Ill-formed vs. Well-formed)*: Responses that do not have syntactic or semantic meaning (i.e., responses with poor sentence structure, responses containing misspelled or unknown words, or responses consisting of incomplete sentences) are *ill-formed* sentences. Responses that have syntactic or semantic meaning are *well-formed* sentences.

> ***TA inquiry*:** Why did you perform divide 2?
> ***Tutor response (Ill-formed)***: jhjgjkgjk (unknown word) / coficient needs to go by dividing (poor sentence structure with misspelled words) / the coefficient (incomplete sentence)
> ***Tutor response (Well-formed)***: because it will cancel out 2.

(ii) *Relevancy (Relevant vs Irrelevant)*: Responses are *relevant* if they can independently reveal some information about the working domain and *irrelevant* if they could belong to any problem-solving domain.

> ***TA inquiry*:** Why did you perform divide 2?
> ***Tutor response (Relevant)***: it will help you solve the equation.
> ***Tutor response (Irrelevant)***: Do as I say / I kind of had a feeling that it may be right

(iii) *Information content* (*Why* vs *what/how*): Responses are *why-informative* if it describes why a solution step or an alternative solution step is correct or incorrect, whereas responses are *what/how-informative* if it describes what solution step to perform or how a solution step is executed.

> ***TA inquiry*:** I thought "subtract 2" applies here. Why am I wrong?
> ***Tutor response (Why)***: It should not be subtract 2 because we have a negative constant with variable term and subtract 2 would make the equation worse. / It should not be subtract 2 because it would make the equation worse.
> ***Tutor response (what/how)***: you need to divide 2 instead. (what) / 3v-2-2 = 3v-4 (how)

(iv) *Intonation (Descriptive vs Reparative)*: Responses are *descriptive* if someone explains their stance and *reparative* if someone acknowledges they made a mistake. Note that a response is not reparative if someone is repairing someone else's mistake.

> ***TA inquiry*:** I thought "subtract 2" applies here. Why am I wrong?
> ***Tutor response (Descriptive)***: Because subtract 2 will not help you get rid of the constant.
> ***Tutor response (Reparative)***: Sorry, it should be subtract 2. I am wrong.

Our first hypothesis towards operationalization is that a knowledge-building or knowledge-telling response must be well-formed, relevant, and either descriptive or reparative. Therefore, any ill-formed, irrelevant tutor responses belong to the "other" class, which is neither knowledge-building nor knowledge-telling, for which the TA must seek a well-formed, relevant response from the tutor. Our second hypothesis is that the information content is what distinguishes knowledge-building responses from knowledge-telling responses. Since why-informative explanation promotes inferences

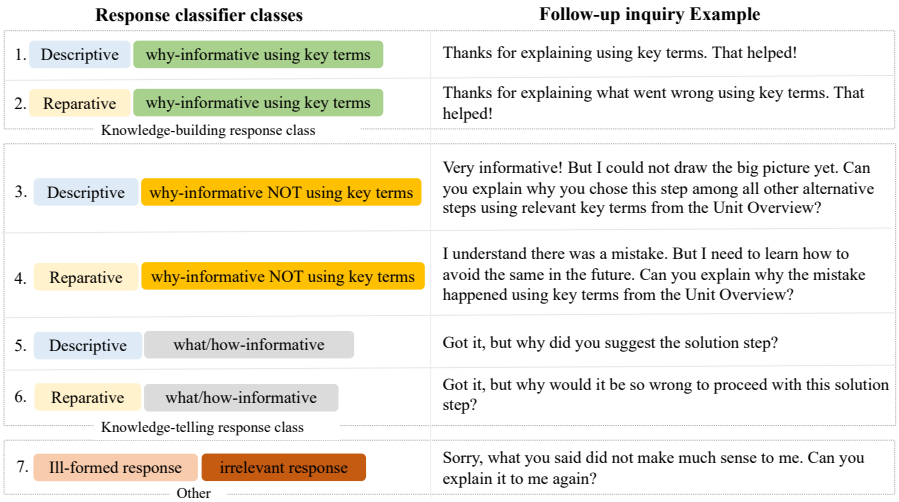

Fig 4: Seven classes of response classifier and the corresponding follow-up inquiry generated by dialog manager. The features in a response based on which the classes are formed is shown using color coded box.

[14, 15] and tutors are likely to realize their gaps by generating inferences [13], our third hypothesis was "descriptive, why-informative responses" or "reparative, why-informative responses" would be knowledge-building responses. Consequently, "descriptive, what/how-informative responses" or "reparative, what/how-informative responses" would be knowledge-telling responses.

To empirically assess the quality of the why-informative response as an indication of knowledge-building, we analyzed 2676 responses from 165 tutors (79 seventh-grade, 51 eighth-grade, and 35 ninth-grade students) from 4 schools during our past studies. We came across instances of why-informative responses that did not have the potential to be a knowledge-building response, e.g. "I performed divide 2 because I want to simplify the equation." or "It should not be subtract 2 because it would make the equation worse." Our analysis revealed that high-gaining tutors with both high and low prior knowledge tend to include domain-dependent key terms (*constant*, *coefficient*, *like terms*, etc.) in their why-informative explanation. An ANCOVA test fitting the post-test score with the count of responses using key terms while controlling the pre-test score indicated that the count of responses using key terms is a reliable predictor of the post-test score; $F(1,162) = 8.2$, $p < 0.01$. Our analysis confirmed findings from another study that explanations using glossary terms facilitate learning [16]. Therefore, we define knowledge-building responses as "descriptive/reparative, why-informative responses using key terms" and knowledge-building responses as "descriptive/reparative, why-informative responses NOT using key terms" and "descriptive/reparative, what/how-informative responses." **Fig 4** shows the seven classes of our response classifier.

Two human coders categorized 2676 responses into our defined seven classes. Based on Cohen's Kappa coefficient, the inter-coder reliability for this coding showed $\kappa =$

0.81. We used an open-source machine learning tool called LightSide [17] to train our learning model from text[1] to identify our defined classes based on the *sentence formation*, *relevancy*, *information content, intonation*, and *presence of key terms.* Next, we describe the proposed scripted dialog manager that generates follow-up inquiries to nudge tutors to generate responses that are *descriptive* or *reparative, contain why-information and* use key terms.

3.3 Dialog manager

We constructed a script to manage what follow-up inquiries to ask based on the tutors' response classes (shown in **Fig 4**). We used an open-source XML script-based dialog manager called Tutalk [18], which uses the output from the response classifier for its decision-making to fashion dialog routes. The dialog manager asks at most three follow-up inquiries until the tutor provides a knowledge-building response, which is operationalized as *descriptive or reparative* responses that contain *why-information* and use domain-dependent key terms. If the knowledge-building response is not provided, the associated transformation (add, subtract, multiply, or divide) is cached. The TA temporarily moves on to the next solution step while informing the tutor that it will again ask the same initial inquiry (e.g., "*Hmm…I am still confused; I will get back to it later. Let's move on for now!*"). The TA again asks the same initial inquiry when the tutor demonstrates or provides negative feedback on the cached transformation.

4 Method

We have conducted an empirical evaluation study to validate the effectiveness of Constructive Tutee Inquiry (CTI). The experiment was conducted as a randomized controlled trial with two conditions: the treatment condition, where SimStudent asks both initial inquiries and follow-up inquiries (i.e., CTI), and the control condition, where SimStudent only asks initial inquiries. In the rest of the paper, we call the treatment and control conditions the CTI and NoCTI conditions, respectively.

A total of 33 middle school students (14 male and 19 female) of 6th-8th grade from various schools participated in the study for monetary compensation. The study sessions were conducted either in-person at our research lab or online using the Zoom video conferencing platform. 20 out of 33 students participated in the in-person sessions, whereas the remaining 13 students participated online. For the online sessions, APLUS was run on a researcher's computer (operated in the research lab), and the participant used APLUS through the Zoom screen-sharing technology.

Students were randomly assigned between conditions ensuring an equal balance of their grades: 17 students (10 in-person and 7 online) for the CTI condition—7 sixth-graders, 7 seventh-graders, 3 eighth-graders—and 16 students (10 in-person and 6 online) for the NoCTI condition—5 sixth-graders, 8 seventh-graders, 3 eighth-graders.

Participants took a pre-test for 30 minutes on the first day of the study. The test consisted of 22 questions. 10 questions were solving-equation questions (with 2 one-step questions, 2 two-step questions, and 6 questions with variables on both sides), and 12 questions were multiple-choice questions used to measure the proficiency of algebra

[1] Our trained model can be found here: <<masked for blind review>>

concepts (1 question to formulate an equation from a word problem; 7 questions to identify variable, constant, positive, negative, and like terms in an equation; and 4 questions to identify the correct state of an equation after performing a transformation on both sides). In the following analysis, we call the 10 solving-equation questions the Procedural Skill Test (PST) and the 12 multiple-choice questions the Conceptual Knowledge Test (CKT).

The highest score any participant could achieve in the overall test, PST and CKT, are 22, 10, and 12, respectively. No partial marks were provided. In the analysis below, the test scores are normalized as the ratio of the score obtained by the participant to the maximum possible score.

A two-tailed unpaired *t*-test confirmed no condition difference on the pre-test. Overall test: M_{CTI} = 0.60±0.18 vs. M_{NoCTI} = 0.63±0.24; $t(30)$ = -0.35, p = 0.73. PST: M_{CTI} = 0.56 ± 0.30 vs. M_{NoCTI} = 0.64±0.30; $t(30)$ = -0.75, p = 0.46. CKT: M_{CTI} = 0.64 ± 0.13 vs. M_{NoCTI} = 0.62±0.22; $t(30)$ = 0.27, p = 0.79.

Immediately after taking the pre-test, all participants watched a 10-minute tutorial video on how to use APLUS. Participants were informed in the video that their goal was to help their synthetic peer (i.e., the TA) pass the quiz. Participants were then free to use APLUS for three days for a total of 2 hours or to complete their goal (i.e., passing the quiz), whichever came first. Upon completion, participants took a 30-minute post-test that was isomorphic to the pre-test.

In the following analysis, we use the learning outcome data (the normalized pre-and post-test scores) along with process data (participants' activities while using APLUS) automatically collected by APLUS, including (but not limited to) interface actions taken by participants, the TA inquiries, and participants' responses.

5 Results

5.1 RQ1: Can tutor responses be accurately classified into knowledge-building and knowledge-telling to drive CTI?

If our operationalized features of knowledge-building response (KBR) truly indicates tutor's reflecting on their understanding, we must see a positive correlation between the frequency of knowledge-building responses (KBR) and learning gain. A regression analysis fitting normalized gain[2] and the frequency of KBR confirmed that KBR is a reliable predictor of learning gain; *normalized gain* = 0.17+0.01*KBR; $F_{KBR}(1,31)=5.08$, $p<.05$. The regression model implies that *providing one more descriptive or reparative responses containing why-information using key terms results in a 1% increase of normalized gain*.

We also conducted the same analysis with knowledge-telling responses (KTRs), which by definition are *descriptive or reparative responses containing why-information NOT using key terms or* containing *what/how information* only. The result showed that KTR was not a predictor of learning gain; $F_{KTR}(1, 31) = 2.66$, $p = 0.11$.

5.2 RQ2: Does CTI facilitate tutor learning?

Fig 5 shows an interaction plot with normalized pre- and post-test score contrasting

[2] $Normalized\ gain = \frac{Normalized\ Post\ score - Normalized\ Pre\ score}{1 - Normalized\ Pre\ score}$

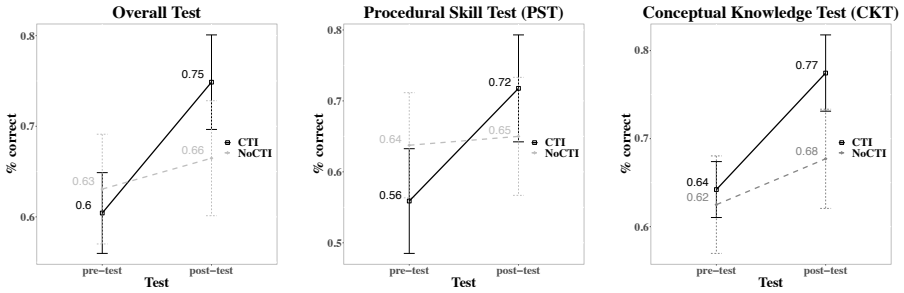

Fig 5: Interaction plot of % correct from pre to post between conditions (CTI vs NoCTI)

CTI and NoCTI conditions. An ANCOVA test with the normalized post-test score as a dependent variable and condition as an independent variable while controlling the normalized pre-test score revealed a main effect of condition, although the effect size was small (M_{CTI} = 0.75±0.22 vs. M_{NoCTI} = 0.66±0. 25; $F_{\text{Condition}}(1, 30) = 5.18$, $p < 0.05$, d = 0.35). We conducted the same analysis for PST (M_{CTI} = 0.72±0.31, M_{NoCTI} = 0.65±0.33; $F_{\text{Condition}}(1, 30) = 3.54$, p = 0.06., d = 0.21) and CKT (M_{CTI} = 0.77±0.18, M_{NoCTI} = 0.68±0.23; $F_{\text{Condition}}(1, 30) = 2.94$, $p = 0.09$, d=0.48). The current data suggests that *CTI tutors had a higher post-test score than the NoCTI tutors, and the discrepancy mostly came from the difference in their proficiency in solving equations*.

5.3 RQ3: Does CTI help tutors learn to engage in knowledge-building?

To understand if CTI tutors learned to engage in more knowledge-building responses, we first conducted a one-way ANOVA on the frequency of knowledge-building responses with condition as a between-subject variable. The result revealed a main effect for condition; M_{CTI} = 19.06±10.21, M_{NoCTI} = 4.19±5.10, $F_{\text{Condition}}(1, 30) = 27.04$, p<0.001. The current data suggests that *CTI tutors generated significantly more knowledge-building responses than NoCTI tutors*.

We further hypothesized that receiving follow-up inquiries helped CTI tutors learn to provide knowledge-building responses (KBR) to initial inquiries over time. To test this hypothesis, we calculated the ratio of KBR to the total responses generated by tutors on every initial-inquiry opportunity they receive. We only considered the first 30 initial inquiries because most of the tutors answered at least 30 initial inquiries before they could complete their goal, i.e., passing the quiz. **Fig 6** shows the chronological change of the ratio of KBR to the total responses provided by tutors on the y-axis and the chronological initial inquiry opportunity on the x-axis. For example, on the 5th initial inquiry, 35% of responses generated by CTI tutors were knowledge-building responses. A two-way ANOVA with the % KBR as the dependent variable and the initial-inquiry opportunity and condition as the independent variable revealed a significant interaction between initial-inquiry opportunity (IIO) and condition: $F_{\text{IIO:condition}}(1, 52) = 10.52$, $p <$ 0.01. *The current data suggests that CTI tutors shifted to provide knowledge-building responses at a higher frequency than NoCTI tutors over time.*

In the CTI condition, SimStudent, the teachable agent, was programmed to ask the same initial inquiry again (as described in section 3.3) if the tutor failed to provide a KBR for a transformation in their previous try. Therefore, one possible reason for CTI

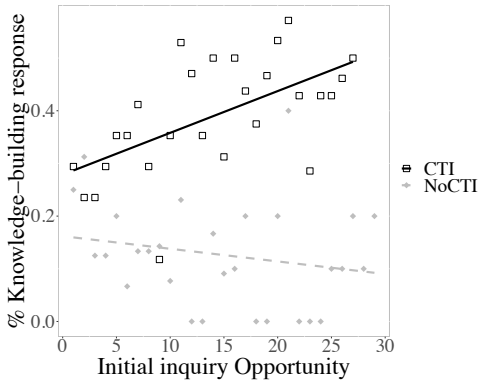

Fig 6: Ratio of knowledge-building responses to the total responses provided by tutors on the subsequent initial inquiry opportunity.

tutors providing more KBR than NoCTI tutors might be that CTI tutors programmatically received more initial inquiries. To our surprise, there was no difference in the average number of initial inquiries tutors received in both conditions; $M_{CTI} = 26.5 \pm 15.15$ vs. $M_{NoCTI} = 30.5 \pm 17.46$, $t(110)=1.30$, $p = 0.20$. *This finding implies that CTI tutors generated more knowledge-building responses than NoCTI tutors even when they received an equal number of initial inquiries.*

6 Discussion

We provided empirical support that an inquisitive teachable agent with constructive inquiries can be built even as a simple application of scripted dialogue, and its effect has been demonstrated. Our proposed Constructive Tutee Inquiry (CTI) helped tutors achieve a higher post-test score than NoCTI tutors, arguably by helping them engage in knowledge-building, like reflecting on their own understanding [4] and recovering from their misconceptions or knowledge gaps [19] while tutoring. The gradual increase of knowledge-building responses on the initial inquiry also suggests that tutors who were incapable of engaging in knowledge-building at the beginning of the study eventually learned to do so over time.

Despite the cogent results on the effectiveness of CTI in the current study, two concerns remain that require scrutiny. First, the small effect size of the current study may suggest room for system improvement. Second, the current data do not show evidence of students' learning gain on the conceptual knowledge test (CKT), even though tutors were prompted to explain solution steps using domain-dependent key terms. One probable reason could be a lack of internal validity for the CKT (i.e., not truly measuring the conceptual knowledge), which requires revision. Alternatively, it could be because CTI only helped tutors acquire shallow skills for solving equations using key terms without the need to deeply understand the meaning of those key terms

Condition	Avg # Knowledge-building responses (KBR)	
	Avg. # Descriptive	Avg. # Reparative
CTI	17.96	1.10
NoCTI	4.19	0

Table 1: KBR broken down into descriptive vs reparative types

or the interconnected relationship among them. Such formulation of shallow skills is argued to only facilitate procedural learning and not conceptual learning [20].

We investigated the nature of the knowledge-building responses (KBR) tutors generated by breaking them down into "descriptive vs. reparative" why-informative responses with key terms, as shown in **Table 1**. The majority of KBR tutors generated were descriptive responses; tutors rarely repaired their mistakes. One possible explanation for this could be that tutors realized their mistake only after exhaustively responding to all the follow-up inquiry turns, and no turns were left for prompting tutors to provide an explanation of why the mistake happened using key terms. Another reason could be tied to a lack of in-depth understanding of the key terms that might have affected their capability to realize mistakes. Then, even if they did realize their mistakes, it was harder to explain why the mistake happened using those terms.

7 Conclusion

We found that guiding tutors towards knowledge-building using Constructive Tutee Inquiry (CTI) eventually helped tutors learn to perform knowledge-building over time without the aid of follow-up inquiries. It acted as implicit training for the tutor to learn to engage in knowledge-building. Furthermore, Answering CTI also helped the tutor enhance their equation-solving skills, which was reflected in their performance. In the literature on learning by teaching, we are the first to develop a teachable agent that facilitates tutor learning with appropriate follow-up inquiries without consulting an expert domain knowledge.

Our analysis casts light upon effective response features that facilitate tutor learning. Our data showed that Constructive Tutee Inquiry was not as effective in tutors' conceptual learning as we expected [10]. Our future interest is to understand the root cause behind this finding and to revise the follow-up inquiries to facilitate both procedural and conceptual learning.

8 References

1. Roscoe, R.D. and M.T.H. Chi, *Understanding tutor learning: Knowledge-building and knowledge-telling in peer tutors' explanations and questions.* Review of Educational Research, 2007. **77**(4): p. 534-574.
2. Chi, M.T.H., et al., *Learning from human tutoring.* Cognitive Science, 2001. **25**: p. 471-533.
3. Graesser, A.C., N.K. Person, and J.P. Magliano, *Collaborative dialogue patterns in naturalistic one-to-one tutoring.* Applied Cognitive Psychology, 1995. **9**(6): p. 495-522.
4. Butler, D.L., *Structuring instruction to promote self-regulated learning by adolescents and adults with learning disabilities.* Exceptionality, 2003. **11**(1): p. 39-60.
5. Hong, H.-Y., et al., *Advancing third graders' reading comprehension through collaborative Knowledge Building: A comparative study in Taiwan.* Computers & Education, 2020. **157**:

p. 103962.
6. Roscoe, R.D. and M. Chi, *Tutor learning: the role of explaining and responding to questions.* Instructional Science, 2008. **36**(4): p. 321-350.
7. Roscoe, R.D., *Self-monitoring and knowledge-building in learning by teaching.* Instructional Science, 2014. **42**(3): p. 327-351.
8. Roscoe, R.D. and M.T.H. Chi, *The influence of the tutee in learning by peer tutoring*, in *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*, K. Forbus, D. Gentner, and T. Regier, Editors. 2004, Erlbaum: Mahwah, NJ. p. 1179-1184.
9. Roscoe, R.D., *Opportunities and barriers for tutor learning: Knowledge-building, metacognition, and motivation*. 2008, University of Pittsburgh.
10. Shahriar, T. and N. Matsuda. *"Can you clarify what you said?": Studying the impact of tutee agents' follow-up questions on tutors' learning*. in *International Conference on Artificial Intelligence in Education*. 2021. Springer.
11. Matsuda, N., et al., *Learning by Teaching SimStudent – An Initial Classroom Baseline Study comparing with Cognitive Tutor*, in *Proceedings of the International Conference on Artificial Intelligence in Education*, G. Biswas and S. Bull, Editors. 2011, Springer: Berlin, Heidelberg. p. 213-221.
12. Li, N., et al., *Integrating representation learning and skill learning in a human-like intelligent agent.* Artificial Intelligence, 2015. **219**: p. 67-91.
13. Roscoe, R.D. and M.T.H. Chi, *Tutor learning: the role of explaining and responding to questions.* Instructional Science, 2008. **36**(4): p. 321-350.
14. Williams, J.J., T. Lombrozo, and B. Rehder, *The hazards of explanation: Overgeneralization in the face of exceptions.* Journal of Experimental Psychology: General, 2013. **142**(4): p. 1006.
15. Rittle-Johnson, B. and A.M. Loehr, *Eliciting explanations: Constraints on when self-explanation aids learning.* Psychonomic bulletin & review, 2017. **24**(5): p. 1501-1510.
16. Aleven, V.A. and K.R. Koedinger, *An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor.* Cognitive science, 2002. **26**(2): p. 147-179.
17. Mayfield, E. and C.P. Rosé, *LightSIDE: Open source machine learning for text*, in *Handbook of automated essay evaluation*. 2013, Routledge. p. 146-157.
18. Carolyn, R., *Tools for authoring a dialogue agent that participates in learning studies.* Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work, 2007. **158**: p. 43.
19. Cohen, J., *Theoretical considerations of peer tutoring.* Psychology in the Schools, 1986. **23**(2): p. 175-186.
20. Nilsson, P., *A framework for investigating qualities of procedural and conceptual knowledge in mathematics—An inferentialist perspective.* Journal for Research in Mathematics Education, 2020. **51**(5): p. 574-599.