Tech Know Learn
DOI 10.1007/s10758-015-9248-6

WORK-IN-PROGRESS

Cognitive Demand of Model Tracing Tutor Tasks: Conceptualizing and Predicting How Deeply Students Engage

Aaron M. Kessler · Mary Kay Stein · Christian D. Schunn

Abstract Model tracing tutors represent a technology designed to mimic key elements of one-on-one human tutoring. We examine the situations in which such supportive computer technologies may devolve into mindless student work with little conceptual understanding or student development. To analyze the support of student intellectual work in the model tracing tutor case, we adapt a cognitive demand framework that has been previously applied with success to teacher-guided mathematics classrooms. This framework is then tested against think-aloud data from students using a model tracing tutor designed to teach proportional reasoning skills in the context of robotics movement planning problems. Individual tutor tasks are coded for designed level of cognitive demand and compared to students' enacted level of cognitive demand. In general, designed levels predicted how students enacted the tasks. However, just as in classrooms, student enactment was often at lower levels of demand than designed. Several contextual design features were associated with this decline. Implications for intelligent tutoring system design and research are discussed.

Keywords Model tracing tutor · Cognitive demand framework · Student task engagement

1 Introduction

A persistent problem of practice in mathematics classrooms is raising the level of student thinking. For reasons that range from poor curricula, to unforgiving bell schedules, to an

This work was funded by grant DRL-1029404 from the National Science Foundation to the second and third authors.

A. M. Kessler (✉) · M. K. Stein · C. D. Schunn
University of Pittsburgh, Pittsburgh, PA, USA
e-mail: aaronmkessler@gmail.com

M. K. Stein
e-mail: mkstein@pitt.edu

C. D. Schunn
e-mail: schunn@pitt.edu

Published online: 19 February 2015

over-reliance on standardized tests, student thinking in the majority of American math classrooms is highly routinized, consisting of memorizing content or reproducing teacher-demonstrated procedures to solve problems (Stigler and Hiebert 2004).

Over the past 20 years, however, there have been some strides. Many students are now at least *exposed* to more complex tasks that challenge them to think, reason, and make sense of new ideas (NRC 2004). This has happened primarily because of standards-setting initiatives—beginning in the late eighties and early nineties—of professional groups such as the National Council of Teachers of Mathematics. After surveying the changing world of work, advances in the discipline, and the challenges of a globalized economy, NCTM presented arguments for why the nature of classroom work needed to become more cognitively complex, developing in our students a host of skills that go beyond the ability to memorize or to repeat back what has already been taught (National Council of Teachers of Mathematics 1989). The standards, in turn, prompted a great deal of teacher professional development along with the design of new curricula that promised to bring twenty-first century skills to our nation's children. Similar work continues in the United States today under the Common Core State Standards.

The mathematics curriculum-based reforms of the nineties were soon followed by research studies that critically examined whether and how the presence of higher-level, curriculum-based tasks actually improved students' learning (e.g., Huntley et al. 2000; Senk and Thompson 2003). In theory, the tasks found in the new curricula should lead to more rigorous forms of thinking and learning. A key finding, however, was that these tasks often changed their character once unleashed in real classrooms settings (Boston and Smith 2009; Stein et al. 1996; Stigler and Hiebert 2004). In particular, teachers often lowered the cognitive demands of tasks by breaking them into smaller subtasks (Smith 2000) and/or focusing on the accuracy of procedures and answers rather than students' thinking and reasoning processes (Henningsen and Stein 1997). Research has also shown that the level of cognitive demand of the *enacted* instructional tasks (not the written tasks) is associated with student gains on measures that target high-level thinking and reasoning. Consistent results over the past 25 years have shown that students learn best when they are in classrooms in which a high-level of cognitive demand is maintained throughout lessons (Boaler and Staples 2008; Hiebert and Wearne 1993; Stein and Lane 1996; Stigler and Hiebert 2004; Tarr et al. 2008).

The above research has been conducted solely with teachers using print-based curricula in conventional classrooms. However, there have also been many computer-assisted efforts to improve student learning in mathematics, efforts that have taken a different form. Model tracing tutors, for example, were developed to assist students to learn how to solve mathematics problems without the teacher. Based on extensive research demonstrating the importance of having both procedural and conceptual knowledge in solving complex tasks (Hiebert 2013; Rittle-Johnson and Alibali 1999; Rittle-Johnson et al. 2001), these tutors focus students on practicing critical aspects they have not yet mastered (Anderson 1996; Ritter et al. 2007). Model tracing tutors aim to reduce the cognitive burden on students, often by "taking over" or heavily guiding them through certain elements in the problem-solving space so that they are free to focus on those elements that still need to be practiced (VanLehn et al. 2000). At the same time, model tracing tutors may have taken over too much of the task, allowing students to practice without attention to the appropriateness of mathematical approaches.

The purpose of the research reported here is to examine the enactment of, and student learning from, computer-based mathematics tasks using a framework that has guided cognitive demand research in mathematics classrooms using print-based materials.

2 Literature Review

2.1 Different Views of Cognitive Demand

Mathematics educators and psychologists have approached the idea of cognitive demand in different ways. Since the call for a shift toward more thinking, reasoning, and problem solving (NCTM 1989), two popular areas of focus of mathematics education researchers has been the nature of students' conceptual understanding (Hiebert and Carpenter 1992) and the relationship between teachers' instruction and students' opportunities to learn (Hiebert and Grouws 2007). Although instruction has been studied in various ways [e.g., through classroom discourse (e.g., O'Connor 2001); teacher questioning (e.g., Boaler and Brodie 2004); intellectual authority (e.g., Wilson and Lloyd 2000)], over the past two decades, the role of challenging and well-structured learning tasks has emerged as a prominent instructional component associated with initiating and sustaining learning processes in mathematics classrooms (Brophy 2000; Baumert et al. 2010; Seidel and Shavelson 2007). Whether referred to as "cognitively activating" (Baumert et al. 2010) or "high cognitive demand" (Stein et al. 1996), there is general consensus regarding what constitutes low- versus high-levels of challenge. Here we use the taxonomy devised by Stein et al. (1996) (see Table 1).

Research has shown, however, that the vast preponderance of tasks used in American classrooms are low level (Grouws et al. 2004; Stigler and Hiebert 2004); as such, a common goal of much teacher professional development has been on learning how to select, set up, and enact tasks at higher levels of cognitive demand.

The model tracing tutor approach, on the other hand, is more consistent with cognitive load theory (Paas et al. 2004; van Merrienboer and Sweller 2005), which postulates that humans have limited cognitive capacity and learning tasks should therefore minimize cognitive demand. In particular, any given learning environment can be analyzed using cognitive load theory to identify three kinds of cognitive loads: intrinsic, the load associated with doing the conceptual task itself; extrinsic, the load associated with operating

Table 1 Levels of cognitive demand ordered from lowest to highest, their definitions, and examples

Level of demand	Definition	Example
Memorization (lowest)	Students recall previously learned facts, rules, formulae, or definitions	What is the Per-unit calculation for computing proportions called?
Procedures without connections	Students follow previously demonstrated algorithms to solve problems without linking them to underlying concepts, meaning or understanding	Divide the number in the left column of the table by the number in the right column to get the unit rate
Procedures with connections	Students follow suggested pathways to solve problems, but do so in a manner that maintains close connections to the underlying mathematical concepts	Use the diagram [a diagram picturing 3 wheel rotations covering 12 cm] to figure how far the robot will travel in 6 rotations
Doing mathematics (highest)	Students are required to solve problems that demand complex nonalgorithmic thinking for which they do not have a predetermined pathway	"Shade 6 small squares in a 4 × 10 rectangle. Using the rectangle, explain how to determine the fractional part that is shaded" (Stein et al. 2009)

within the external task environment (e.g., finding information, operating an interface); and germane, the load associated with reflecting on task performance and learning to improve task performance. Here, cognitive capacity means how much can be handled at once by the human brain and cognitive demand refers to aspects of the problem situation (i.e., cognitive load) that could overtax that capacity.

Model tracing tutors are often designed to minimize all three forms of cognitive load such that the learner is not overwhelmed with a total load that is beyond their cognitive capacity. To minimize intrinsic load, larger tasks are typically divided into steps that each require a response, and the components of larger tasks are practiced to the point of automaticity. Sometimes conceptually irrelevant components of the larger task are solved automatically for the student (e.g., via a calculator). The tutor's computer interface is also often used to show intermediate values or an overall situation model, thereby allowing the external environment to serve as a replacement for internal memory demands (Jang and Schunn 2014; Jang et al. 2011). To minimize extrinsic load, interfaces are made highly regular across problems so students learn where to expect to find information and enter responses. To minimize germane load, immediate feedback is presented whenever a step is done incorrectly, which includes hints about what was done incorrectly and what action would be correct. All of this suggests a very different approach to raising students' level of thinking and reasoning; instead of focusing on the whole task and the processes that students are encouraged to use to solve it, cognitive load theory divides the task into manageable pieces until mastery is reached and the pieces can be combined.

At base, the mathematics education view of cognitive demand and cognitive load theory agree that students should learn how to think, reason, and problem solve; the mathematics education view would also endorse the idea of "off-loading" procedural elements (as in the case of a student who uses a graphing calculator to plot a line graph so as to be "freed up" to focus on characteristics of the graph and how they relate to parameters in the equation). However, the two theories differ in regards to where the emphases have been placed, with model tracing tutors focusing on lowering cognitive demand to building efficient procedural performance/conceptual understanding and mathematics educators focusing on assisting teachers' capacities to set up and support student thinking during high-demand tasks.

2.2 The Advantages of a Task-Based Framework

The tasks with which students become engaged are important because they determine not only what substance students learn, but also how they come to think about, develop, use, and make sense of that substance. Tasks provide signals not only about *what* is important to learn, but also *how* one should learn it and what "counts" as important intellectual activity. The classification of mathematical instructional tasks into various levels of cognitive demand using the Task Analysis Guide (TAG; Stein et al. 2009) coupled with the tracking of tasks from the pages of textbooks to their actual enactment in classrooms using the Mathematics Task Framework (MTF; Stein et al. 1996) have proven to be useful tools for both research and practice. For researchers, these tools have allowed us to document how the cognitive demand of instructional tasks can change as they progress through phases: first, as curricular or instructional materials (print), second as they are set up by the teacher in the classroom (set up), and, finally, as implemented by students during the lesson (enactment) (Boston and Smith 2009; Henningsen and Stein 1997; Stein et al. 1996; Stein and Kaufman 2010). This pattern of shifts relates to work by Bosch et al. (2006) on didactical transpositions and by Venezky (1992) on the "curricular chain" that includes the

desired, prescribed, delivered, and received curriculum. Practitioners, teachers and teacher educators find that the TAG and MTF provide a language for them to talk about things that often happened in classrooms but for which they had no way of expressing. Here we explore the utility of the TAG and MTF for examining student learning with model tracing tutors: will it also help explain the depth of student thinking during tutor-based learning tasks?

Following research in mathematics education, we will also use instructional tasks as our unit of analysis. However, instead of three phases (print, set up, and enactment), in tutoring environments, the first two phases are combined into one phase; there is no teacher in the environment to "set up" the task. Thus, the task is delivered to the student exactly as intended by the designers with no opportunities for a drop in demand until the task is actually enacted by the students. However, there is still the opportunity for students to decline cognitive demand during enactment, and we hypothesize that features of the task environment can predict the circumstances under which students are more likely to decline cognitive demand.

In sum, we examine whether the TAG and MTF framework (based in one notion of cognitive demand) can be usefully applied to analyzing learning in model-tracing tutors, whose design was rooted in a different notion of cognitive demand. Can the TAG and MTF provide useful insights into the level and kind of thinking in which students engage while working on computer-based tasks, especially the model tracing tutors that are designed to be teacher-less? If, indeed, the decline in cognitive demand in mathematics education studies was often rooted in teacher actions, might these model tracing tutor environments be better positioned to maintain the level of thinking and reasoning of students as they engage with the tasks?

More specifically this research asks three questions related to the mathematics task framework and students' enactment of an intelligent tutor built within a popular model tracing tutor environment:

1. Does the cognitive demand of the designed task predict the cognitive demand when the students enact the task in a model tracing tutor context?
2. Is the students' level of thinking while enacting model tracing tutor tasks associated with learning?
3. What, if any, factors are associated with the maintenance of cognitive demand from design to student enactment in a model-tracing tutor?

Theoretically, the study tests the generalizability of the MTF and TAG to less teacher-centric learning environments. Pragmatically, the study can provide guidance on how to improve intelligent tutor design by revealing which design features tend to produce problematic student learning behaviors.

3 Methods

3.1 Participants

This work was conducted at a small suburban public school district over the course of 3 weeks totaling 16 h of student engagement with the tutor system. Given the tutor's use of robotics as a motivating context (see below), the project was presented to students, grades 6–8, enrolled in the districts robotics club as a way to gain experience with robots, which were used as part of the district's competitive robotics team, and to build mathematics

skills. In total, the twenty-seven students in the robotics club were contacted for participation in the program. Of the twenty-seven contacted, nine showed interest and six males and one female (M = 12.1, SD = .7, range 11–13) ultimately completing the program. Each of the seven students completed 8 units in the system consisting of approximately 10–18 tutor pages per unit.

3.2 The Tutoring Environment

This project analyzed an Intelligent Tutoring System (ITS) called the Robot Algebra Project (RAP). The RAP was developed within the model tracing tutor framework used to create many different systems (algebra, geometry, and physics), enacted by large numbers of students throughout the world (VanLehn et al. 2000, 2005; Ritter et al. 2007; Koedinger and Corbett 2006). Studying an ITS of this type provides relevance to a broad class of systems in the tutoring world. Hallmarks of these tutors are: (1) they track skills through students' actions; (2) feedback is tailored to the particular skills they appear to be missing; and (3) practice focuses on topics for skills which the students have not developed (Ma et al. 2014).

The RAP design team, made up of researchers with decades of experience in tutor design, robotics education, and mathematics education, designed a system to improve students' proportional reasoning in the context of rich robotics problems. The studied version of RAP had been iteratively improved across 10 months of testing in different classroom and after-school contexts, building upon a computational framework (CTAT; http://ctat.pact.cs.cmu.edu) that allowed for rapid tutor development. That is, the studied tutor was likely representative of the popular model tracing tutors deployed throughout the world, rather than a poorly designed system produced with little development effort or by designers not well trained in the design task.

Proportional reasoning is a key proficiency that students need to develop during the middle school years (Lamon 2007). Although students typically learn to execute an algorithm for solving proportional reasoning problems quite easily (cross multiplication), they struggle with the underlying concept of what is meant by a proportional relationship (the simultaneous coordination of changes in two interrelated quantities) and the development of flexible use of multiple representations and strategies for solving problems that involve proportionality (Lobato et al. 2010; Harel et al. 1994; Thompson and Saldanha 2003).

Ideally, the designed tutor would function independent of other strong instructional resources, such as mathematics teachers, so that the learning could take place in technology classrooms or various informal learning environments (e.g., home schooling, summer camps, after school clubs). Building upon prior computer-based materials that successfully taught proportional reasoning using LEGO robotics activities (Silk et al. 2011), the design team built a system that integrates the use of a LEGO robot and NXT software with a model tracing tutor to introduce both robotics and proportional reasoning to students ranging in grades four through eight. LEGO robotics provides a good fit to proportional reasoning because most aspects of motion programming in these simple robots involve proportional relationships (e.g., linear distance, linear speed, turn angle, and turn speed are all linear functions of wheel size and motor rotations; turn angle and turn speed is also a linear function of axle width). The system also provides a physical model in which reasoning can happen. This physicality appears to support students' ability to reason about mathematical relationships by making more salient the critical aspects of the robot that produce proportional relationships (Silk and Schunn 2011; Silk et al. 2011; Liu and Schunn

2014), similar to other kinds of embodiment support for learning (Alibali and Nathan 2012).

The tutoring environment is accessed via an Internet-connected computer, and the activities also depend upon using a LEGO robot and the NXT software that is used to program the robot. Students work individually with the tutor. Although some learning contexts require students to share robots for cost reasons, the studied context provided one robot per student. Figure 1 shows a typical page in RAP that requires students to utilize the robot to gain information necessary to complete a tutor task. Note that the computer interface provides many of the on-screen supports to minimize student cognitive load in the mathematics tasks (i.e., step-by-step guidance on the larger task; required intermediate calculations; feedback and hints as needed). Not included on this page, but present on all problem pages, is a skillometer, which shows students their progress on various component proportional reasoning skills being assessed by the tutor.

The entire RAP tutor consists of nine instructional units (see Fig. 2): one unit introducing the student to robot basics (e.g., how to install programs on the robot and run the programs), three short units that teach measurement of different aspects of robotics motion, a challenge problem involving many measurement tasks, three longer units on proportional relationships in motion, and then a challenge problem involving motion programming. We selected the unit on proportional distance as the focal point for this study because it is

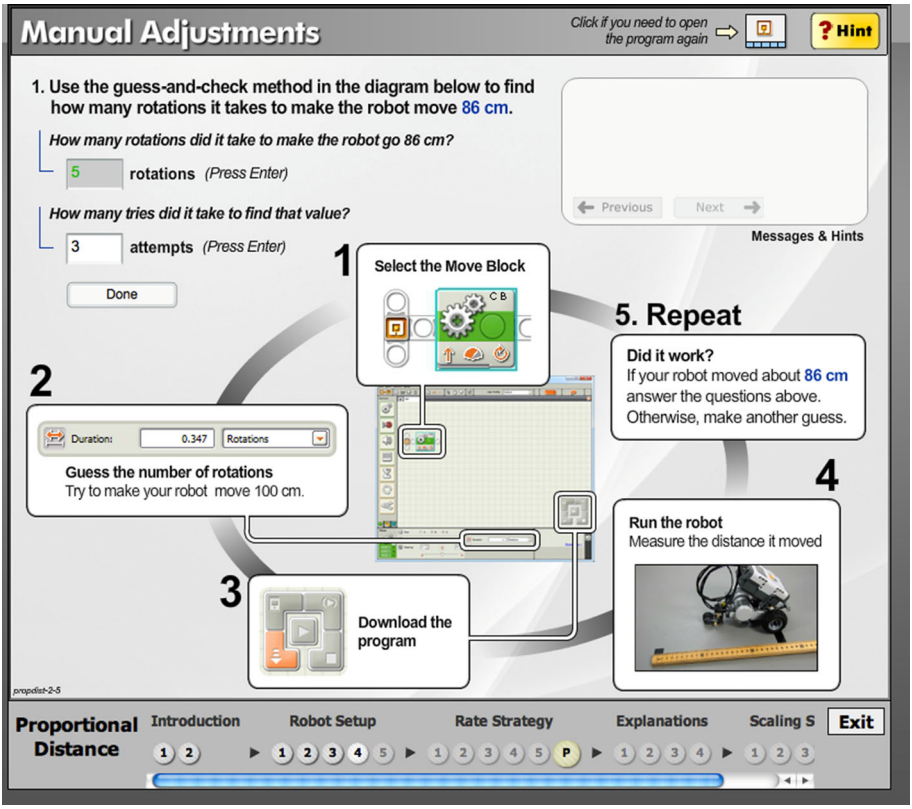

Fig. 1 A RAP example page

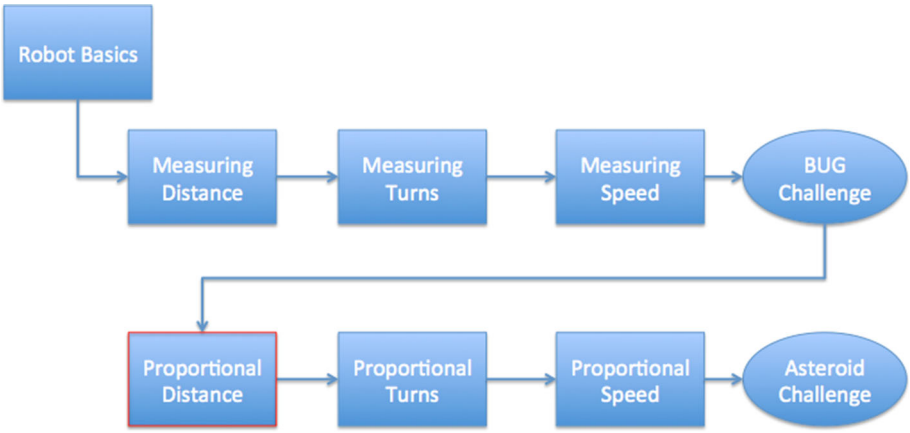

Fig. 2 The sequence of units in RAP, and the unit selected for analysis (proportional distance)

where the key mathematics concepts underlying proportional reasoning are first introduced (i.e., where the most conceptual work related the proportional reasoning is done).

Successful proportional reasoning involves being able to flexibly apply different strategies to solve proportional problems according to the problem situation (Lobato et al. 2010). One key strategy choice is the distinction between unit-rate and scaling factor strategies. Consider the problem: if one robot moves 10 cm with 30 motor rotations, then how many rotations are required to move 20 cm? A unit rate strategy would involve determining that the robot moves 1 cm per 3 motor rations, and thus 20 cm would require $3 \times 20 = 60$ rotations. By contrast, a scaling strategy would involve seeing that the new distance is twice the old distance, and thus the new distance would require $2 \times 30 = 60$ rotations. In this problem, both strategies are easy. But depending on the numbers involved, one strategy can be much easier than the other strategy, and thus it is useful to be able to flexibly apply both strategies. Across the instructional units, students are positioned to learn both unit-rate and scaling strategies.

3.3 Measures

3.3.1 Designed Level of Cognitive Demand

To answer the first research question, the designed level of cognitive demand for each task was established. First, we defined the boundaries of a task as the images, tables, feedback, and answer options associated with each tutor *question*. Therefore, a single tutor page could consist of up to 7 individual tasks. The analysis of designed cognitive demand level was completed before students began work on the tutor. Screen capture images of each tutor question and all possible combinations of answers and feedback for that question were used to code for the designed level of cognitive demand using the same levels as described in Table 1 (see Fig. 3 for an example). Next, the questions on the screen shots of the tutor pages were coded by a second researcher to establish an inter-rater reliability of 95 % agreement (Miles and Huberman 1994).

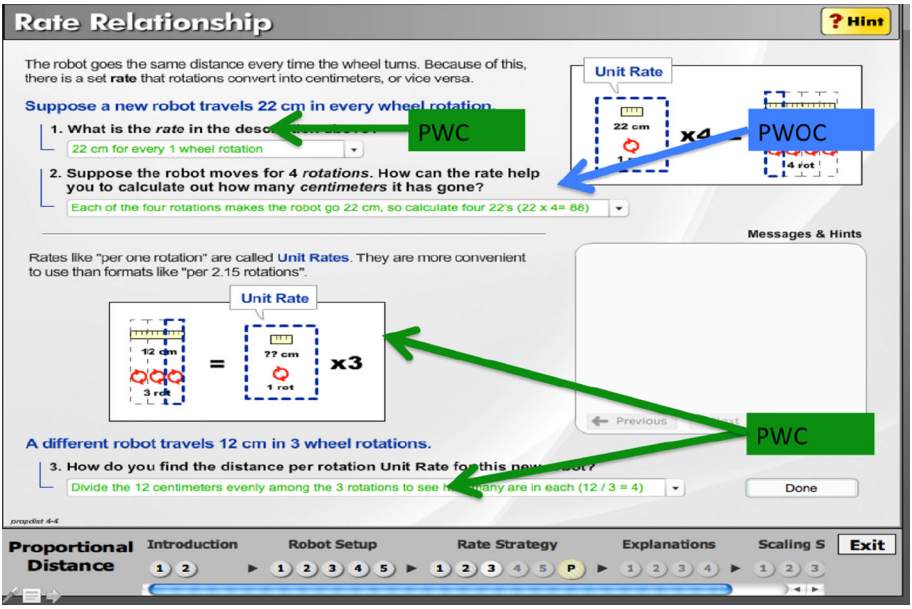

Fig. 3 Example cognitive demand coding of tasks within the tutor

3.3.2 Enacted Cognitive Demand

All three research questions required coding the enacted cognitive demand of each task. Due to the low number of Memorization and Doing Mathematics tasks (4 and <1 % respectively) that were identified at the *designed* level, coding for the *enacted* level of cognitive demand focused only on questions that were given a designed code of the Procedures With Connects (PWC) and Procedures Without Connections (PWOC). Within the cognitive demand literature, PWC are considered high level tasks and PWOC are considered low level tasks. Thus, dropping Memorization and Doing Mathematics from analysis does not limit the study's main goal of predicting high and low cognitive demand of student enactment.

Students' cognition as they individually worked on the RAP tasks (the enacted cognitive demand) was measured using student think-aloud protocols (Ericsson and Simon 1993; Carmel et al. 1992). This allowed the research team to examine student thinking in more depth than the log data analysis strategies typically used by other tutor researchers (VanLehn et al. 2005). Additionally, in comparison to test or survey-based methods, the think aloud method provides direct access to students' reasoning *during* learning. Finally, it produces many data points for each participant (i.e., the total number of data points can be large enough to support quantitative analyses as well) providing data amenable to qualitative analysis. A limitation of using a think aloud method is a reliance on a relatively small number of participants. As such, it is important to select a variety of students who would likely engage with the system, which we have done in both gender and age.

All screen movements and interactions with the tutor and robot were video taped along with each student's verbalization of what they were thinking about as they worked through the problems. As they talked through their ideas, the students were also asked to read out

loud what was on the screen. The videos were transcribed, and analyses were based on the transcripts together with the video of behaviors in the tutor windows to provide context.

Before coding for cognitive demand, the researcher coded for three types of talk (Reading, Task Talk, and Robot-related Talk). Reading was defined as students reading verbatim off the tutor screen. Task talk was student talk that revolved around the task presented to them and included the thinking they were doing around each proportional reasoning task. Robot-related talk was talk surrounding their efforts to make the robot function (away from the tutor itself) and was outside the scope of the task presented in the tutor. For this analysis, only the Task Talk was analyzed because it represented the level of cognitive thinking exhibited by students during their enactment of tutor tasks.

Specifically, students' thinking surrounding each PWC and PWOC question on the tutor screen was coded as Memorization, Procedures Without Connections, Procedures With Connections or Doing Mathematics, using the same code definitions that were used for coding of the screen captures. Two raters coding 20 % of the total transcripts had high agreement (89 % exact match) in coding for enacted cognitive demand. In order to obtain an "average" enactment code for each question to compare against design codes, the codes were averaged across the seven students. Again, these calculations were carried out only on PWC and PWOC questions due to the low number of Memorization and Doing Mathematics demand questions.

3.3.3 Student Learning Outcomes

Students were given an independent paper and pencil assessment of proportional reasoning: a 13-question assessment of proportional reasoning adapted from Weaver and Junker (2004), previously shown to have high reliability and validity for use with middle school students. There were two alternative forms administered in a counter-balanced fashion across the participants. The pre-assessment was administered prior to the beginning of the Robot Basics Unit. The post assessment was administered after the completion of the Asteroid Challenge. We assigned two separate codes to each of the student responses on the Weaver & Junker assessment. The first code represented whether or not the student had answered the question correctly regardless of the process, strategy, or amount of work shown. The second code represented the strategy that was used by the student to answer the question, which was only possible when students showed enough work in order for the researcher to identify the strategy.

We created five dichotomous variables to represent strategy use. The first simply coded for an incorrect strategy. Students scored a "0" if they used a wrong strategy or failed to produce a correct answer. The remaining four dichotomous indicators were scored "1" for the presence of the unit rate strategy, scale factor strategy, cross multiplication or another strategy.

3.3.4 Factors Associated with Maintenance or Decline of Task Demand

To determine whether specific contextual factors could predict whether questions designed at a high level of cognitive demand ($n = 29$) tended to be enacted at the high level of demand (or decline), both the screen shots and talk-aloud data were coded using an adapted set of factors from previous research (Doyle 1983; Stein et al. 1996). The factors, definitions, and mean percentage of high demand questions that received each code are presented in Table 2.

Table 2 Possible factors associated with maintenance or decline of cognitive demand

Factor	Definition	% of high demand questions
Task feedback	Type of tutor feedback causes a drop in the demand of the original task by taking over and doing the thinking for the student	64
Task relevance	Task is explicitly related to the overall purpose of the tutor challenge	55
Task set up: steps to solve	The steps to solve the task are explicitly given as part of the page (not hints)	38
Task simplified interface	The task has a unique interface; it is not a replication of previous tasks formats	37
Task sequence computation	Task drops to demanding only computation over a sequence of questions	28
Task accountability	A skillometer was present for the question (because skill level as shown on the skillometer determined whether students could move on to new sections of the tutor)	24
Task thinking not valued	Task only expects students to provide quick and correct solution to the problem	24

4 Results and Discussion

4.1 Level of Cognitive Demand as Designed

Given the interactive tutoring system context, memorization tasks were quite rare. Further, given the focus on closed-ended responses that could easily be processed by the tutoring system, "doing mathematics" tasks were also rare. Therefore, the bulk of the tasks were PWC and PWOC, with roughly an equal distribution (see Table 3). These results verified that the designed level of cognitive demand for this tutor contained both high demand, meaning-oriented tasks shown in research to improve mathematics learning, and low demand tasks, established with increasing skill proficiency in solving math problems.

4.2 Enacted Levels of Cognitive Demand

Table 4 presents the mean percentage of question responses coded at each level of cognitive demand. Similar to the designed level, PWC and PWOC are the dominant enacted levels of cognitive demand in this context. However, at the enacted level, PWOC becomes much more common.

4.3 Relationship Between Designed and Enacted Levels of Cognitive Demand

To address the first research question, each designed question code was compared with the average level of enactment across all the students for each specific task. Given the relatively small number of participants, here we are testing the robustness of effect of designed cognitive demand level on enacted cognitive demand level by examining the robustness of the effects across the PWOC and PWC ($n = 51$) tasks in the tutor. The results show a significant difference for students' level of enactment at the PWC level for questions at different design levels (i.e., designed at the PWC or PWOC levels) ($t = 3.004$, $df = 44.97$,

Table 3 Distribution of designed cognitive demand

Cognitive demand level	# of questions
No code	5 (9 %)
Memorization	2 (3 %)
Procedures without connections	22 (38 %)
Procedures with connections	29 (50 %)
Doing mathematics	0 (0 %)

Table 4 Percentage of observed enacted cognitive demand for tasks designed at the PWC or PWOC levels

Cognitive demand level	% of question responses
No code	5
Memorization	2
Procedures without connections	80
Procedures with connections	13
Doing mathematics	0

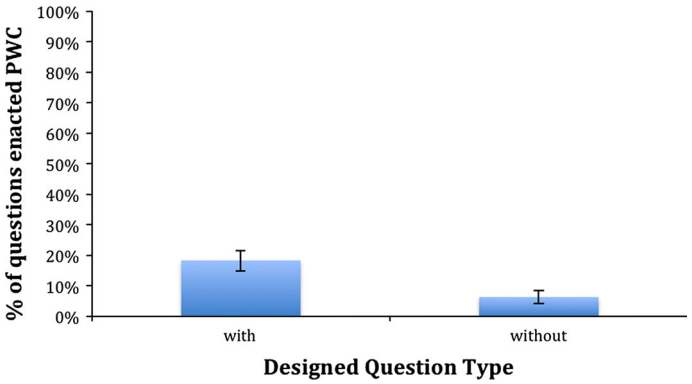

Fig. 4 The percentage of tasks enacted at procedures with connections for tasks designed at the procedures with and without connections levels

$p = .004$, $d = .85$). This large and statistically significant difference suggests that the designed level of the task was a general predictor of the level at which students enacted a question.

Figure 4 also reveals a general trend towards decline of cognitive demand: students very often enacted questions designed to be PWC as PWOC. By contrast, students almost never enacted questions at higher levels of demand than how they were designed (i.e., a PWOC task was rarely enacted at a PWC level). This preponderance towards decline rather than increase of cognitive demand from design to enactment mirrors what is found in typical mathematics classrooms (Henningsen and Stein 1997).

4.4 Learning and Strategy Shift

In order to answer the second research question, we began by focusing on the changes in performance on the pre/post Junker/Weaver measure of proportional reasoning. There was

no overall gain from pre (n = 7, $M = 14.0$, SD = 3.0) to post (n = 7, $M = 13.9$, SD = 2.7, d = .05). The lack of gain may be ceiling effect of this particular population: 83 % correct on the Junker/Weaver pre-test. However, some of the participants began at lower levels and they showed no gains. Further, pre-post changes assessed in two other settings with less mathematically sophisticated learners using the same tutoring system also found no gains.

Although no significant difference was seen in their accuracy scores, there were some differences in strategy use from pre to post assessments. In the first case study, presented below, we further unpack the nature of these strategy shifts as seen in both the think-aloud and the assessments.

4.5 Factors Associated with Maintenance and Decline of Cognitive Demand

Addressing the third research question, we conducted a stepwise multivariate regression at the question level (n = 29 tasks designed at the PWC level) to test which of the 7 possible factors coded from the think-alouds and screen captures independently predict decline of demand (i.e., proportion of students enacting each task at the PWOC level). We found two statistically significant predictors (which are also statistically significant as separate correlations): task accountability (i.e., the presence of a skillometer or not) and quality of task feedback in the hints (see Table 5). Note that task accountability prevented decline whereas task feedback increased decline, as one would expect. As shown by the β values, both effects were quite large.

4.6 Case Studies

To examine in further depth the nature of student thinking and learning with the tutor, we present two case-study students using both qualitative and quantitative data. The students were selected because they were average for the sample (i.e., they are reading at or above grade level, and they have an interest in robotics) and they were particularly articulate about how they made choices. Further, one was selected to be relatively strong and one to be relatively weak in prior mathematical ability.

The first case reveals the ways in which a student's low-level engagement with the tutor materials caused a shift in her strategy use while resulting in no gains in terms of her problem accuracy. The second case illustrates the ways in which accountability and feedback can impact the maintenance of the cognitive demand of tasks in the tutor.

4.6.1 Emma

As a 7th grader, Emma's pre assessment score on the Weaver & Junker evaluation of proportional reasoning was a 53 %, suggesting that she had plenty of room for

Table 5 Multiple regression predicting decline of demand at the question level

Predictor	B	Std. error	β	t	*p* value
Task accountability	−.277	.078	−.67	−3.57	.002
Task feedback	.184	.071	.49	2.60	.016

improvement through engagement with the tutor. During the pre-assessment she also showed a lack of correct strategy use, employing a proportional strategy to solve the problems on only 4 of 17 questions. Despite this lack of proportional reasoning at pre-test, she breezed through the early measurement portions of the tutor finishing the BUG Challenge before any other student.

From analysis of her think-aloud, it became clear that her biggest concern was figuring out how to solve the problems required for her to progress through the tutor and not taking the time to understand what was going on or any of the contextual framing that surrounds the tutor tasks. This procedural focus was demonstrated in her skipping the introductory pages to every section. These pages were designed to frame the tasks and give real world meaning to the material. She chose, instead, to move quickly past these pages without reading any of the materials or looking at the diagrams or pictures.

During her think-aloud across 47 questions, she was assigned a cognitive demand level of PWOC 41 times. Most of her tutor-talk revolved around developing a way of solving the problems and then duplicating that process moving forward in her work. The exchange below, between the tutor, researcher, and Emma, suggests she was duplicating what she had done to correctly solve a previous problem rather than thinking about what the problem was asking and whether or not that strategy was correct for a new task. Figure 5 highlights the question she is talking about while showing the other questions on the tutor page.

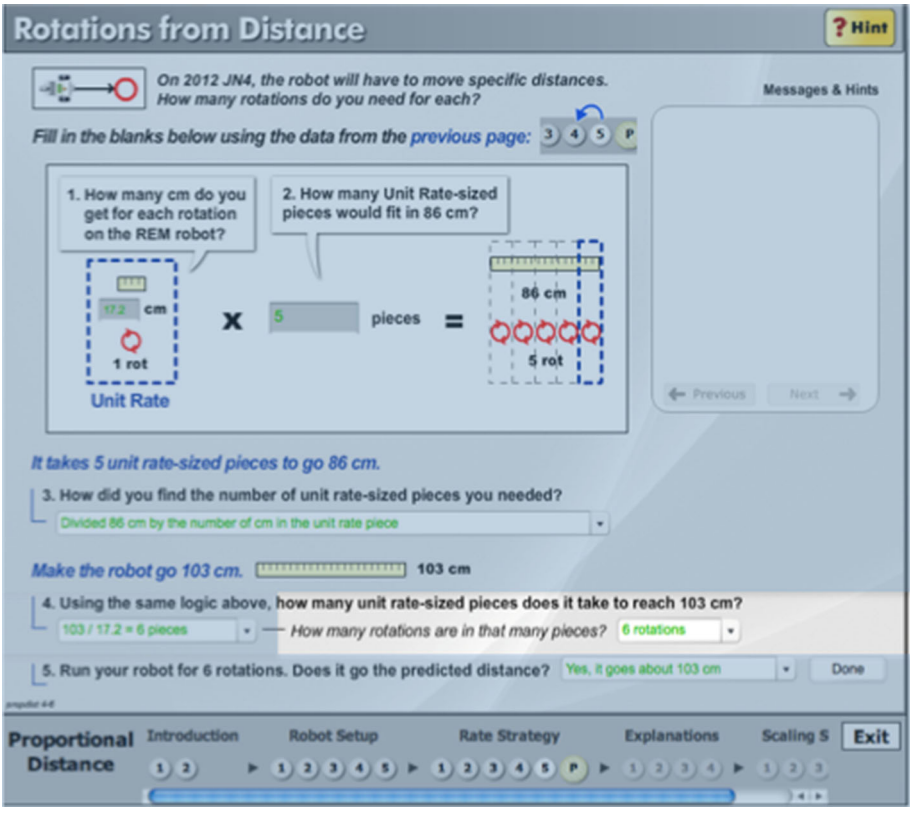

Fig. 5 Highlighted task talk question

Task Question: How many rotations are in that many pieces?
Emma: Pauses and says nothing
Researcher: "And what are you doing right now?"
Emma: "Um, I am…I think that I should divided these 2 but. How many rotations are in that many pieces? On the last page I did 7."
Researcher: "What are you looking at on the last page?"
Emma: "I was looking at how I did it before"

Rather than thinking about how unit rate could be applied to the new problem, Emma instead references back to the method she used to solve the questions on the previous page to answer the new task, without much thought as to whether the process is appropriate for this new problem.

Because she experienced success during the unit rate pages, the desire to carry forward those same procedures led to her become very frustrated when the tutor asked her to switch strategies and use scaling. This frustration was best demonstrated in the second scaling strategy task. In this exchange, the tutor is asking her to unpack the process of finding the scale factor that exists between two similar units. Figure 6 highlights that question she is talking about while showing the other questions on the tutor page.

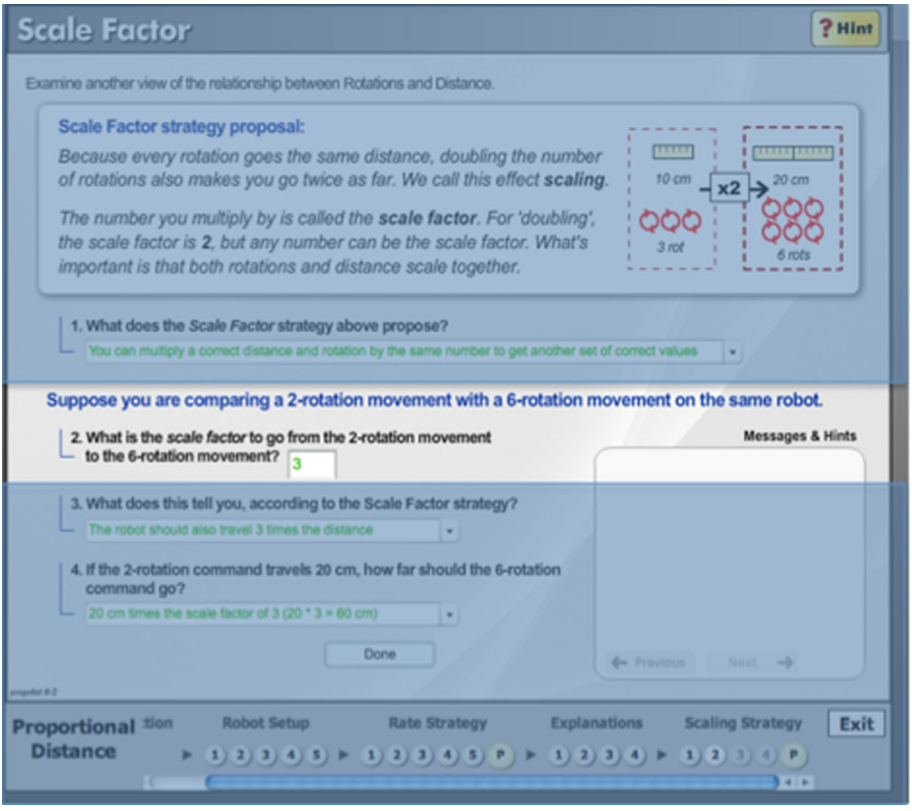

Fig. 6 Highlighted task talk question 2

Task Question: So what is the scale factor to go from the two-rotation movement to the six-rotation movement?
Emma: "Um, so I am doing 20 divided by 10 to find the scale factor or cm which is 2. So lets go with just 2. Oh wait…"
Tutor Hint: Supposed you are comparing a two-rotation movement with a six-rotation movement. What is the scale factor to go from the two-rotation movement to the six-rotation movement?
Emma: "I need a hint"
Tutor Hint: Compare the problem with the report above. In the report the scale factor is 2 because the number of rotations is going from 2 to 4 and so did the distance. So apply…
Emma: "oh wait I get it. So I just have to find the scale factor. Constant rate, sorry constant rate, which is 20 divided by 6 which is 3.3."

This confusion between unit rate and scale factor continues for almost another 3 min before the tutor finally describes to her exactly what to plug into the provided box. Her confusion regarding the difference between thinking about the proportion within the units (scale factor) versus between units (unit rate) suggests a lack of conceptual grounding for thinking about proportionality. Instead of thinking and reasoning her way through the scale factor problems, she misapplied the unit-rate strategy, most likely because of her past successes using this for solving previous tutor problems

Emma's affinity to the strategy with which she had success continued on the post assessment where she utilized the unit rate strategy on 10 of the 17 questions. This dramatic increase in use of the unit rate strategy, although encouraging in that she was using a proportional strategy (whereas she did so on a very limited basis on the pre assessment), had little impact on her ability to actually get the correct answer, as her average score actually decreased from the pre-assessment to a 47 %. This suggests that she was unable to build the procedural and conceptual knowledge necessary to fully implement the unit rate strategy even though it appeared to be strategy with which she was most comfortable.

4.6.2 Tim

As a 6th grader in the study, Tim's success on the pre assessment was much more representative of the other five students in the study, completing 88 % of the problems correctly. Although this score left some room for improvement, it is clear that he had an understanding of solving proportional reasoning problems before the project began. Unlike the other five students' pre-assessment, who showed a mix of methods in solving the problems, Tim showed a preference for the scale factor strategy, answering 7 of the 17 questions in this way. Tim, like Emma, moved quickly through the early measurement units of the tutor and finished the BUG challenge at the same time as most of the other students, all within about 10 min of each other.

In the following transcript, Tim figures out that if he asks for hints his skill level does not increase, requiring him to solve more problems. Following Tim's realization, he focuses carefully on his actions and what he needs to do on the next page of the tutor so that his skill level will increase.

Tim: "I finished those three all green and got the going the distance achievement. (the skillometer is still yellow) Awww, looks like I have one more."
(The tutor moves to next page)

Tim: "Um…ok what did I do last time? I think I did 75 times, so, shift 8, times 90 enter. (the answer is wrong and Tim asks the tutor for a hint) Oh! 84 times .75. Cause that is the distance and this is the distance. 63. Done!"
(The skillometer does not change and the tutor loads another problem)
Tim: "Oh, I am guessing if I ask for a hint it doesn't move me any further."
(After getting another problem correct and the skillometer not changing)
Tim: "Ok I need to get it right this time!"
(Tim; slows down and takes much more time with the next problem then he had with the previous two)
Tim: "Ok, 6, 27, 18…So 6–18 would be, come on, come on, YES. Ok I know both of these so I do 18 divided by 6 equals 3. 27 times 3 would be 81."
(Tim gets the problem correct and the skillometer increases)
Tim: "Oh yes I am getting the hang of this."
(Tim goes on to take his time with the next three problems getting each correct)

In this example, Tim spends much more time thinking about what he should enter into the computer. The accountability associated with the skillometer focuses Tim's attention to the importance of thinking about the problem. This accountability proved to be a significant factor across all students.

5 General Discussion

5.1 Theoretical Implications

Research and development surrounding ITS has primarily focused on building student capacity to learn procedural problem-solving (Ma et al. 2014; VanLehn 2011). This has resulted in a set of systems that focus on students' ability to solve problems correctly, building procedural fluency, but with little attention to conceptual understanding of the topic.

The RAP tutor system, on the other hand, introduced proportional reasoning knowledge through tasks (questions) at different cognitive demand levels (PWC and PWOC) (Stein et al. 1996) over multiple pages of a model tracing tutor. The introduction of high-demand tasks to the RAP system suggested the potential utility of analyzing students' engagement with the RAP using a framework outside the traditional tutor literature, specifically the TAG and MTF (Stein et al. 1996, 2009). The results show that the TAG and MTF in combination with a talk aloud procedure can be used to evaluate the cognitive demand level of both designed and enacted tutor tasks. In using these frameworks, which are well documented in the mathematics literature, we have developed a way to frame and study tutor tasks' impact on the way students are engaging with the system, specific factors that can be associated with the decline or maintenance of cognitive demand, and how this impacts students' learning.

5.2 Pedagogical Implications

At first glance, some of the reported results seem to contradict each other (e.g., a significant relationship between cognitive demand levels at the design and enactment phases, but a drop in demand during critical pages). However, using the various pieces of data collected during this study we can begin to unpack these puzzling results.

Past research suggests that high-demand tasks that are enacted at high levels result in greater learning gains on high-demand assessments (Stein and Lane 1996). The quantitative results of level of cognitive demand at the design versus enacted phases are important in that they suggest that designing higher level demand tasks into a computer-based curriculum can have an impact on the level at which students enact tasks. These results suggest that by attending to the cognitive demand of tasks during the tutor design process we can increase the level of student-enacted demand. In other words, by posing higher demand tasks, the designers are providing students with more opportunity to use procedures in a way that connects to conceptual understanding (as contrasted with tutors that are designed with predominantly low-demand, PWOC, tasks). This makes the design work and associated cost with developing these higher-level tasks a value-added endeavor. However, the discrepancy between the quantitative and qualitative data suggests that designing high-demand tasks, while necessary to encourage students' conceptual reasoning, is not sufficient in moving all students to deep thinking about the tasks.

Much like the work from the print-based curricula research (Boston and Smith 2009; Stein et al. 1996; Stigler and Hiebert 2004), our findings suggests that the designed demand of tasks is often lowered when actually enacted by students. From the think-aloud data we know that the design of the task itself is not always (or even usually) sufficient to guarantee enactment at a high level. While there were some tasks designed to be high level and students tended to enact them at a high level, there were also tasks that were designed as high level but were commonly enacted at a low level. Here, students were falling back on the procedural knowledge learned and required to quickly solve problems and were not taking the time or energy necessary to develop the deep and cross-cutting conceptual understanding of proportional reasoning.

From the analysis of factors associated with the maintenance of demand during the enactment phase, it appears that the lack of a skillometer on these pages is one possible reason why students did not think and reason at a high level. Signaling to students that they are accountable, like in Tim's case, for the conceptual understanding of proportions, a tutor could help to maintain the desired level of demand for particular tasks. Additionally, as students were given feedback and hints for these tasks (limiting the germane cognitive load), the text often encouraged students to simplify and proceduralize the tasks rather than maintaining a focus on meaning. Together, the lack of accountability combined with feedback that does not maintain a focus on meaning too-often resulted in students' level of demand declining while enacting tasks designed at a high demand.

Finally, although this study shows no changes in students' ability to correctly solve proportional reasoning problems, it does demonstrate such a system's ability to shift some students' strategy use. Students like Emma and Tim, who show this shift, point out a possible strength of the system. The tutor enabled these students to shift at least part of their thinking and moved both of them in a direction of at least using a conceptually appropriate proportional strategy in the post assessments.

6 Conclusion

This study makes three important contributions. First, it demonstrates how using a think-aloud protocol in conjunction with the TAG and MTF framework allows different kinds and levels of student thinking to be identified and described in a model tracing tutor context. This method provides researchers with a more multi-faceted exploration of student

learning by evaluating enacted levels of cognitive demand in these unique and quickly evolving environments.

Second, if the incorporation of high demand tasks into the RAP represents a new development and direction for the widely used model tracing tutors, this study suggests that researchers need to better understand the impact such tasks will have on students' engagement and learning. Past research suggests that more complex conceptual knowledge tasks will be more difficult for students to enact at high levels and that lower-level enactments have not been associated with improved learning outcomes (Stein and Lane 1996). More research is needed to completely understand how to design and support students' engagement with tasks so they are taken up at a high level of demand. This study shows the importance of working towards such a goal by identifying two important factors (accountability and task feedback) that should be considered in model tracing tutor design principals moving forward. Future work should also consider such factors as gamification; reading level analysis; how changing the context, setup, and style of questions could encourage students into higher or lower levels of task demand enactment; and how altering the feedback text and supports to be less procedural might impact enactment of demand.

Finally, this work calls attention to the decline of cognitive demand level during conceptually challenging tasks. Based on informal observations, development meetings, and user feedback from this project, one possible solution comes from introducing a human instructor into the system (Kessler et al. 2014). This instructor mediation of tasks could provide just enough of a press on students to maintain higher-level demand enactment ultimately producing the deeper understanding of both procedural and conceptual knowledge that the tutor sets out to teach.

References

Alibali, M. W., & Nathan, M. J. (2012). Embodiment in mathematics teaching and learning: Evidence from learners' and teachers' gestures. *Journal of the Learning Sciences, 21*(2), 247–286.

Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American Psychologist, 51*(4), 355–365.

Baumert, J., Kunter, M., Blum, W., Voss, T., Jordan, A., Klusmann, U., et al. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal, 47*(1), 133–180.

Boaler, J., & Brodie, K. (2004). The importance of depth and breadth in the analysis of teaching: A framework for analyzing teacher questions. In *Proceedings of the 26th meeting of the North American chapter of the international group for the psychology of mathematics education*. Toronto, Ontario, Canada.

Boaler, J., & Staples, M. (2008). Creating mathematical futures through an equitable teaching approach: The case of Railside School. *The Teachers College Record, 110*(3), 608–645.

Bosch, M., Chevallard, Y., & Gascón, J. (2006). Science or magic? The use of models and theories in didactics of mathematics. In M. Bosch (Eds.), *Proceedings of the fourth congress of the European Society for research in mathematics education* (pp. 1254–1263). Barcelona.

Boston, M. D., & Smith, M. S. (2009). Transforming secondary mathematics teaching: Increasing the cognitive demands of instructional tasks used in teachers' classrooms. *Journal for Research in Mathematics Education*, *40*(2),119–156.

Brophy, J. (2000). *Teaching*. Brussels: International Academy of Education.

Carmel, E., Crawford, S., & Chen, H. (1992). Browsing in hypertext: A cognitive study. *IEEE Transactions on Systems, Man, and Cybernetics, 22*, 865–884.

Doyle, W. (1983). Academic work. *Review of Educational Research, 53*, 159–199.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (revised ed.). Cambridge, MA: MIT Press.

Grouws, D., Smith, M. S., & Sztajn, P. (2004). The preparation an detaching practices of U.S. mathematics teachers: 1990–2000 mathematics assessments of the *National Assessment of Educational Progress; Results and interpretations* (pp. 221–269). Reston, VA: National Council of Teachers of Mathematics.

Harel, G., Behr, M., Lesh, R., & Post, T. (1994). Invariance of ratio: The case of children's anticipatory scheme for constancy of taste. *Journal for Research in Mathematics Education, 25*, 324–345.

Henningsen, M., & Stein, M. K. (1997). Mathematical tasks and student cognition: Classroom-based factors that support and inhibit high-level mathematical thinking and reasoning. *Journal for Research in Mathematics Education*, 524–549.

Hiebert, J. (Ed.). (2013). *Conceptual and procedural knowledge: The case of mathematics*. London: Routledge.

Hiebert, J., & Carpenter, T. (1992). Learning and teaching with understanding. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 65–97). New York: Macmillan.

Hiebert, J., & Grouws, D. (2007). The effects of classroom mathematics teaching on students' learning. In F. Lester (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 371–404). Charlotte, NC: Information Age.

Hiebert, J., & Wearne, D. (1993). Instructional tasks, classroom discourse, and students' learning in second-grade arithmetic. *American Educational Research Journal, 30*(2), 393–425.

Huntley, M. A., Rasmussen, R. S., Villarubi, J., Sangtong, J., & Fey, J. T. (2000). Effects of standards-based mathematics education: A study of the Core-plus mathematics project algebra and functions strand. *Journal for Research in Mathematics Education, 31*(3), 328–361.

Jang, J., & Schunn, C. D. (2014). A framework for unpacking cognitive benefits of distributed complex visual displays. *Journal of Experimental Psychology Applied 20*(3), 260–269.

Jang, J., Schunn, C.D., & Nokes, T. J. (2011). Spatially distributed instructions improve learning outcomes and efficiency. *Journal of Educational Psychology*, *103*(1), 60–72.

Kessler, A., Boston, M., & Stein, M. K. (2014). Conceptualizing Teacher's Practices in Supporting Students' Mathematical Learning in Computer-Directed Learning Environments. Report published in the *proceedings of the 11th International Conference of the Learning Sciences*, Boulder, CO.

Koedinger, K., & Corbett, A. (2006). Cognitive tutors: Technology bringing learning science to the classroom. In K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences*. Cambridge, MA: Cambridge University Press.

Lamon, S. J. (2007). Rational numbers and proportional reasoning: Toward a theoretical framework for research. *Second Handbook of Research on Mathematics Teaching and Learning, 1*, 629–667.

Liu, A., & Schunn, C. D. (2014). Applying math onto mechanism: Investigating the relationship between mechanistic and mathematical understanding. Paper presented at the 36th annual meeting of the cognitive science society. Quebec City, Canada.

Lobato, J., Ellis, A. B., Charles, R., & Zbiek, R. (2010). *Developing essential understanding of ratios, proportions & proportional reasoning: Grades 6–8*. Reston, VA: National Council of Teachers of Mathematics.

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology, 106*(4), 901–918.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis*. Thousand Oaks, CA: Sage.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: National Council of Teachers of Mathematics.

National Research Council. (2004). *On evaluating curricular effectiveness: Judging the quality of K-12 mathematics evaluations* (Committee for a Review of the Evaluation Data on the Effectiveness of NSF-Supported and Commercially Generated Mathematics Curriculum Materials). Washington, DC: The National Academies Press.

O'Connor, M. C. (2001). Can any fraction be turned into a decimal? A case study of a mathematical group discussion. *Educational Studies in Mathematics, 46*, 143–185.

Paas, F. G. W. C., Renkl, A., & Sweller, J. (2004). Cognitive load theory: Instructional implications of the interaction between information structures and cognitive architecture. *Instructional Science, 32*, 1–8.

Ritter, S., Anderson, J. R., Keodinger, K. R., & Corbett, A. (2007). Cognitive tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review, 14*(2), 249–255.

Rittle-Johnson, B., & Alibali, M. W. (1999). Conceptual and procedural knowledge of mathematics: Does one lead to the other? *Journal of Educational Psychology, 91*(1), 175.

Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology, 93*(2), 346.

Seidel, T., & Shavelson, R. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*(4), 454–499.

Senk, S. L., & Thompspn, D. R. (Eds.). (2003). *Standards-based school mathematics curricula: What are they? What do students learn?*. Mahwah, NJ: Erlbaum.

Silk, E. M., Higashi, R., & Schunn, C. D. (2011). Resources for robot competition success: Assessing math use in grade-school-level engineering design. In *American Society for Engineering Education*. American Society for Engineering Education.

Silk, E. M., & Schunn, C. D. (2011). Calculational versus mechanistic mathematics in propelling the development of physical knowledge. Paper presented at the 41st annual meeting of The Jean Piaget Society, Berkeley, CA.

Smith, M. S. (2000). Balancing old and new: An experienced middle school teacher's learning in the context of mathematics instructional reform. *The Elementary School Journal*, *100*(4), 351–375.

Stein, M. K., Grover, B. W., & Henningsen, M. (1996). Building student capacity for mathematical thinking and reasoning: An analysis of mathematical tasks used in reform classrooms. *American Educational Research Journal, 33*(2), 455–488.

Stein, M. K., & Lane, S. (1996). Instructional tasks and the development of student capacity to think and reason: An analysis of the relationship between teaching and learning in a reform mathematics project. *Educational Research and Evaluation*, *2*(1), 50–80.

Stein, M. K., & Kaufman, J. H. (2010). Selecting and supporting the use of mathematics curricula at scale. *American Educational Research Journal*, *47*(3), 663–693.

Stein, M. K., Smith, M. S., Henningsen, M. A., & Silver, E. A. (2009). *Implementing standards-based mathematics instruction: A casebook for professional development*. New York, NY: Teachers College Press, Columbia University.

Stigler, J. W., & Hiebert, J. (2004). Improving mathematics teaching. *Educational Leadership, 61*(5), 12–17.

Tarr, J. E., Reys, R. E., Reys, B. J., Chávez, Ó., Shih, J., & Osterlind, S. J. (2008). The impact of middle-grades mathematics curricula and the classroom learning environment on student achievement. *Journal for Research in Mathematics Education*, *39*, 247–280.

Thompson, P. W., & Saldanha, L. (2003). Fractions and multiplicative reasoning. In J. Kilpatrick, W. G. Martin, & D. Schifter (Eds.), *A research companion to principles and standards for school mathematics*. Reston, VA: National Council of Teachers of Mathematics.

van Merrienboer, J. J. G., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review, 17*(2), 147–177.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*, 197–221.

VanLehn, K., Freedman, R., Jordan, P., Murray, C., Osan, R., Ringenberg, M., et al. (2000). Fading and deepening: The next steps for Andes and other model-tracing tutors. In G. Guathier, C. Frasson, & K. VanLehn (Eds.), *Intelligent tutoring systems* (pp. 474–483). Berlin: Springer.

VanLehn, K., Lynch, C., Schulze, K., Shapiro, J. A., Shelby, R., Taylor, L., et al. (2005). The Andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence in Education, 15*(3), 147–204.

Venezky, R. (1992). Textbooks in school and society. In P. Jackson. (Ed.), *Handbook of research on curriculum* (pp. 436–462). New York: Macmillan.

Weaver, R., & Junker, B. W. (2004). Model specification for cognitive assessment of proportional reasoning (No. 777). Department of Statistics Technical Report.

Wilson, M., & Lloyd, G. (2000). The challenge to share mathematical authority with students: High school teachers' experiences reforming classroom roles and activities through curriculum implementation. *Journal of Curriculum and Supervision, 15*, 146–169.