Rus and Ştefănescu *Smart Learning Environments* (2016) 3:2
DOI 10.1186/s40561-016-0025-3

Smart Learning Environments
a SpringerOpen Journal

RESEARCH Open Access

Non-intrusive assessment of learners prior knowledge in dialogue-based intelligent tutoring systems

Vasile Rus* and Dan Ştefănescu

*Correspondence:
vrus@memphis.edu
Department of Computer Science, The University of Memphis, 323 Dunn Hall, 38152 Memphis, TN, USA

Abstract

Goal and Scope: This article describes a study whose goal was to assess students' prior knowledge level with respect to a target domain based solely on characteristics of the natural language interaction between students and conversational Intelligent Tutoring Systems (ITSs). We report results on data collected from two conversational ITSs: a micro-adaptive-only ITS and a fully-adaptive (micro- and macro-adaptive) ITS. These two ITSs are in fact different versions of the state-of-the-art conversational ITS DeepTutor (http://www.deeptutor.org).

Approach and Results: Our models rely on both dialogue and session interaction features including time on task, student generated content features (e.g., vocabulary size or domain specific concept use), and pedagogy-related features (e.g., level of scaffolding measured as number of hints). Linear regression models were explored based on these features in order to predict students' knowledge level, as measured with a multiple-choice pre-test, and yielded in the best cases an $r = 0.949$ and adjusted r-square $= 0.833$. We discuss implications of our findings for the development of future ITSs.

Keywords: Learner assessment, Dialogue-based intelligent tutoring systems, Educational technologies

Introduction

Assessment is a key element in education in general and in Intelligent Tutoring Systems (ITSs; (Rus et al. 2013)) in particular because fully adaptive tutoring presupposes accurate assessment (Chi et al. 2001; Woolf 2008). Indeed, a necessary step towards instruction adaptation is assessing students' knowledge state such that appropriate instructional tasks (macro-adaptation) are selected and appropriate scaffolding is offered while students are working on a task (micro-adaptation or within-task adaptation).

We focus in this article on assessing students' prior knowledge in dialogue-based ITSs based on characteristics of the tutorial dialogue interaction between students and such systems. Assessing students' other states, e.g. affective state, that are important for learning and therefore important to further adapt instruction to each individual learner is beyond the scope of this work.

When students start interacting with an ITS, their prior knowledge with respect to the target domain is typically assessed using a multiple choice pre-test although other forms

of assessment such as open answer problem solving are sometimes used. The pre-test serves two purposes: enabling macro-adaptation in ITSs, i.e. the selection of appropriate instructional tasks for a student based on student's knowledge state before the tutoring session starts, and, when paired with a post-test, establishing a baseline from which the student progress is gauged by computing learning gains (post- minus pre-test score). This widely used pre-test/post-test experimental framework is often necessary in order to infer whether the treatment was effective relative to the control.

While the role of a pre-test is important for assessing students' prior knowledge, there are several challenges with having a pre-test. First, a pre-test (as well as the paired post-test) takes up a non-trivial amount of time. This is particularly true for experiments consisting of only one session in which case the pre-test and post-test may take up to half the time of the full experiment. For instance, a 2-h experiment could be broken down into three parts: 30 min for pre-test, 1 h of actual interaction with an ITS, and 30 min for post-test. Altogether, in this particular case the pre-test and post-test take 1 h which is half the time of the whole experiment.

More worryingly is the fact that in such experiments the pre-test may have a tiring effect on students. By the time students reach the post-test many of them will be so tired they will underperform even if they learned something during the actual training, thus, jeopardizing the whole experiment. For instance, in one of our experiments about 30 % of the subjects simply randomly picked one of the choices for the multiple-choice questions in the post-test without even reading the question. We observed this behavior by analyzing the time students took to pick their choice after they were shown a question on screen. About a third of the students took on average less than 5 s per question which is not even enough to read the text of the question. By comparison, the same students took on average 36 s to respond to similar questions in the pre-test. By eliminating the pre-test in the above illustrative experiment, we can reduce the overall experimental time to 1 h and 30 min, thus reducing tiring effects. By eliminating both the pre-test and post-test, we can further reduce the total experiment time.

Additionally, many times there is a disconnect between the pre- and post-test questions and the actual learning tasks and process. To overcome this challenge, Shute and Ventura (2013) argue for a shift towards emphasizing performance-based assessment which is about evaluating students' skills and knowledge while applying them in authentic contexts. For instance, reading instructions in a role-playing game allows assessing students' reading comprehension skills (Shute and Ventura 2013). Using explicit tests in such contexts would interfere with the main task and are therefore not recommended. They advocate for the use of stealth assessment while students engage in a particular activity. Like in stealth assessment, we advocate here for non-intrusive assessment during problem solving in dialogue-based ITSs. To this end, the goal of our work presented here was to investigate to what degree we can automatically infer students' knowledge level directly from their performance while engaging in problem solving with the help of an ITS.

Eliminating the need for learners to go through a standard pre-test and a post-test saves time for more training, eliminates tiring effects and testing anxieties, and ultimately provides a more accurate picture of students' capabilities as the assessment is conducted in context, i.e. while they solve problems in our case. In particular, we investigate how well we can predict students' prior knowledge, as measured by a standard multiple-choice pre-test, based on characteristics of the tutorial dialogue interaction with the hope that if the

predictions are close enough we can do without the pre-test in the future. We are also interested in finding out the minimum tutorial dialogue interaction that would yield an accurate estimate of students' prior knowledge.

We would like to emphasize that we are not arguing for a complete elimination of explicit assessments such as multiple-choice tests which have their own advantages for learning such as testing effects (the memory retrieval processes activated during testing benefit long-term memory of the target material; (Roediger and Karpicke 2006)). Rather, we propose to investigate to what extent we can measure students' knowledge level from interaction characteristics such that, when needed, we can employ this kind of non-intrusive assessment.

We conducted our research on data collected from an experiment with high-school students using the state-of-the-art conversational computer tutor DeepTutor (Rus et al. 2013). As mentioned, our goal was to find interaction features that are good predictors of students' pre-test scores and to create prediction models that would be as useful as the multiple choice pre-tests in measuring students' prior knowledge. The best model we found can predict students' prior knowledge, as measured by a summative pre-test, with $r = 0.949$ and adjusted r-square $= 0.833$. We also determined the minimum dialogue length which is necessary to be able to make the best predictions.

The remainder of the article is organized as follows: Section "Related work" briefly discusses previous relevant work while Section "DeepTutor: a state-of-the-art dialogue-based intelligent tutoring system" presents a brief overview of the computer tutor that provided the context for our experimental analysis. The following section decribes the approach. The data is presented in the next section which is followed by the "Experiments and results" section offering details about the various prediction models and the results we obtained from these models. The article ends with a section on conclusions and further work.

Related work

The most directly relevant previous work to ours is by Lintean et al. (2012) who studied the problem of inferring students' prior knowledge based on prior knowledge activation (PKA) paragraphs elicited from students. PKAs were generated by students as part of a meta-cognitive training program. Lintean and colleagues employed a myriad of methods to predict students' prior knowledge including comparing the student PKA paragraphs to expert-generated paragraphs or to a taxonomy of concepts related to the target domain, which in their case was biology. Students' prior knowledge level or mental model were modeled as a set of three categories: low mental model, medium mental model, and high mental model. There are significant differences between our work and theirs. First, we deal with dialogues as opposed to explicitly elicited prior knowledge paragraphs. Second, we do not have access to a taxonomy of concepts against which we can compare students' contributions. Third, we model students' prior knowledge using scores obtained on a multiple-choice pre-test.

Predicting students' learning and satisfaction is another area of research directly relevant to ours. Among these, we mention the work of Forbes-Riley and Litman (2006) who used three types of features to predict learning and user satisfaction: system specific, tutoring specific, and user-affect-related. They employed the whole training session as unit of analysis, which is different from our own analysis because we use instructional

task, i.e. a Physics problem in our case, as the unit of analysis. Our unit of analysis serves better our purpose of finding out the minimum number of leading instructional tasks to accurately assess students' knowledge level. Furthermore, their work was in the context of a spoken dialogue system while in our case we focus on a chat-based/typed-text-based conversational ITS. Another difference between our work and theirs is their focusing on user satisfaction and learning while we focus on identifying students' knowledge level.

Williams and D'Mello (2010) worked on predicting the quality of student answers (as error-ridden, vague, partially-correct or correct) to human tutor questions, based on dictionary-based dialogue features previously shown to be good detectors of cognitive processes (cf. (Williams and D'Mello 2010)). To extract these features, they used LIWC (Linguistic Inquiry and Word Count; (Pennebaker et al. 2001)), a text analysis software program that calculates the degree to which people use various categories of words across a wide array of texts genres. They reported that pronouns (e.g. I, they, those) and discrepant terms (e.g. should, could, would) are good predictors of the conceptual quality of student responses.

Yoo and Kim (2012) worked on predicting the project performance of students and student groups based on stepwise regression analysis on dialogue features in Online Q&A discussions. To extract dialogue features they made use of LIWC and speech acts, which are semantic categories such as *Greetings* or *Questions* that indicate speakers' intentions (Moldovan et al. 2011). Yoo and Kim found that the degree of information provided by students and how early they start to discuss before the deadline, are two important factors explaining project grades. A similar research was conducted by Romero and colleagues (Romero et al. 2013) who also included (social) network related features. Their statistical analysis showed that the best predictors related to students' dialogue are the number of contributions (messages), number of words, and the average score of the messages.

In our work presented here, we use some of the features described by the above researchers, such as session length or dialogue turn length, and other novel features such as information content.

DeepTutor: a state-of-the-art dialogue-based intelligent tutoring system

The work described in this article has been conducted in the context of the state-of-the-art intelligent tutoring system DeepTutor (http://www.deeptutor.org). To better understand this context, we offer in this section an overview of intelligent tutoring systems in general and of DeepTutor in particular.

Intelligent tutoring systems

One-on-one human tutoring is one of the most effective solutions to instruction and learning that has attracted the attention of many for decades. Encouraged by the effectiveness of one-on-one human tutoring (Bloom 1984), computer tutors such as DeepTutor that mimic human tutors have been successfully built with the hope that a computer tutor could be available to every child with access to a computer (Rus et al. 2013).

How effective are state-of-the-art ITSs at inducing learning gains in students?

An extensive review of tutoring research by VanLehn (2011) showed that computer tutors are as effective as human tutors. VanLehn reviewed studies published between 1975

and 2010 that compared the effectiveness of human tutoring, computer-based tutoring, and no tutoring. The conclusion was that the effectiveness of human tutoring is not as high as it was originally believed (effect size d = 2.0) but much lower (d = 0.79). The effectiveness of computer tutors (d = 0.78) was found to be as high as the effectiveness of human tutors. So, there is something about the one-on-one connection that is critical, whether the student communicates with humans or computers. Graesser et al. (1995) argued that the remedial part of tutorial interaction in which tutor and tutee collaboratively improve an initial answer to a problem is the primary advantage of tutoring over classroom instruction. Chi et al. (2004) advanced a related hypothesis: tutoring enhances students' capacity to reflect iteratively and actively on domain knowledge. Furthermore, one-on-one instruction has the advantage of engaging most students' attention and interest as opposed to other forms of instruction such as lecturing/monologue in which the student may or may not choose to pay attention (VanLehn et al. 2007).

Dialogue-based intelligent tutoring systems

Intelligent Tutoring Systems (ITSs) with conversational dialogue form a special category of ITSs. The development of conversational ITSs such as DeepTutor is driven by explanation-based constructivist theories of learning and the collaborative constructive activities that occur during human tutoring (Rus et al. 2013). Conversational ITSs have several advantages over other types of ITSs. They encourage deep learning as students are required to explain their reasoning and reflect on their basic approach to solving a problem. Such conceptual reasoning is more challenging and beneficial than mechanical application of mathematical formulas (Hestenes et al. 1992). Furthermore, conversational ITSs have the potential of giving students the opportunity to learn the language of scientists, an important goal in science literacy. A student associated with a more shallow understanding of a science topic uses more informal language as opposed to more scientific accounts (Mohan et al. 2009).

DeepTutor

DeepTutor is a state-of-the-art conversational ITS that is intended to increase the effectiveness of conversational ITSs by promoting deep learning of complex science topics through a combination of advanced domain modeling methods, deep language and discourse processing algorithms, and advanced tutorial strategies. DeepTutor is the first ITS based on the framework of Learning Progressions (LPs; (Corcoran et al. 2009)). LPs, which were developed by the science education research community, can be viewed as incrementally more sophisticated ways to think about an idea that emerge naturally while students move toward expert-level understanding of the idea. DeepTutor is an effective ITS: a recent experiment showed that DeepTutor is as effective as human tutors (Rus et al. 2014) yielding effect sizes comparable to the effectiveness of human tutors as reported by VanLehn (2011).

DeepTutor currently targets the domain of conceptual Newtonian Physics but it is designed with scalability in mind (cross-topic, cross-domain). DeepTutor has been developed as a web service and a first prototype is fully accessible through a browser from any Internet-connected device, including regular desktop computers and mobile devices such as tablets, thus moving us closer to the vision of providing

cost-effective and tailored instruction to every learner, child or adult, anywhere, anytime.

The spin-off project of AuthorTutor (http://www.authortutor.org) aims at efficiently porting DeepTutor-like ITSs to new domains by investigating well-defined principles and processes as well as developing software tools that would enable experts to efficiently author conversational computer tutors across STEM disciplines. Another authoring tool, called SEMILAR (derived from SEMantic simILARity toolkit; (Rus et al. 2013)), is being developed as well to assist with authoring algorithms for deep natural language processing of student input in conversational ITSs. More information about the SEMILAR toolkit is available at http://www.semanticsimilarity.org.

It is beyond the scope of this article to describe all the novel aspects of DeepTutor or related projects. Instead, we present next the general instructional framework in DeepTutor with an emphasis on macro- and micro-adaptation which is important to know in order to better understand the data analyses presented in this article.

We would like to just mention that DeepTutor proposed major improvements in core ITSs tasks: modeling the task domain, tracking students' knowledge states, selecting appropriate learning trajectories, and the feedback mechanisms. Advances in these core tutoring tasks will move state-of-the-art ITSs closer to implementing fully adaptive tutoring which means tailoring instruction to each individual student.

The DeepTutor instructional framework

During a typical tutorial session with DeepTutor, the system challenges students to solve a number of problems that are carefully selected by the system in order to optimize student learning. A snapshot of the learner view in DeepTutor is shown in Fig. 1. When working on a particular problem, students are first asked to provide a solution that must include a justification based on concepts and principles of the target domain, which is Newtonian Physics in the case of our study presented here. The student answer is automatically analyzed by comparing it with an ideal answer, provided by an expert apriori, using advanced natural language processing methods (Rus et al. 2013). If there are missing steps in the student solution, the system encourages the student to discover and articulate them with the help of progressively informative hints in the form of questions (e.g., *Which of Newton's laws is relevant to this situation and why?*, *What does Newton's second law say?*). Furthermore, if the student articulates misconceptions the DeepTutor system immediately corrects them. That is, DeepTutor encourages students to self-explain their solution and only offers help when the student is floundering.

All other things equal, low knowledge students will most likely struggle to provide solid self-explanations and therefore most likely to omit important steps in the solution and articulate misconceptions which would lead to more scaffolding dialogue moves in terms of hints and correcting misconceptions, respectively, on the part of the computer tutor. High knowledge students would need less scaffolding and therefore the corresponding dialogues should be shorter. That is, each dialogue between the system and a student has a unique signature or dialogue interaction fingerprint which we exploit in our work here in order to infer students' prior knowledge.

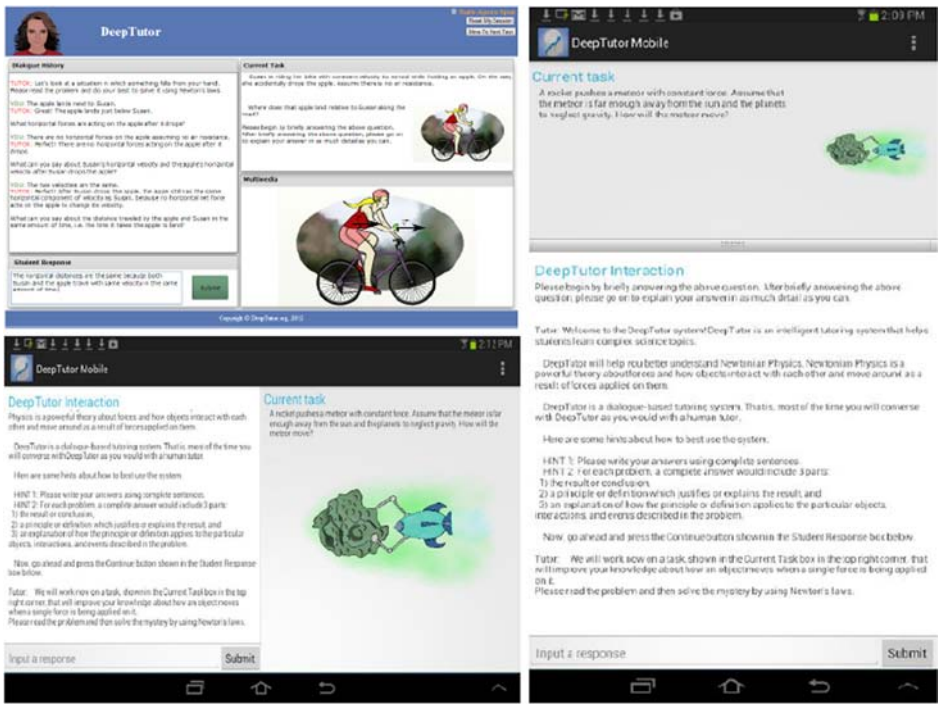

Fig. 1 Snapshot of the browser-based DeepTutor interface. Snapshot of the DeepTutor interface as seen by the learner. The top-left image shows the browser-based interface showing a Physics problem on the top right pane and the Dialogue history on the left pane. The Multimedia box (bottom-right pane) synchronizes with the dialogue such that information identified in the dialogue is represented graphically in the multimedia box, e.g. showing velocity and force vectors graphically once mentioned in the dialogue. The bottom (horizontal view) and right (vertical view) images show the app-based interface that students see if using an Android or iOS app to access DeepTutor

Macro- and micro-adaptivity in DeepTutor: the 3-loop instructional framework

The behavior of DeepTutor can be described using three major loops: *the task loop*, *the solution-step loop*, and *the hint loop*. This framework was inspired from VanLehn's two-loop characterization of tutoring systems (VanLehn et al. 2007). According to VanLehn, ITSs can be described in broad terms as running two loops: the outer loop, which selects the next task to work on, and the inner loop, which manages the student-system interaction while the student works on a particular task. The outer loop corresponds to our task loop while the inner loop corresponds to both the solution-step and hint loops.

We added a third loop, the hint loop as shown in Fig. 2, to better manage the within-task (or micro-level) adaptivity in DeepTutor. Such a separation of the solution-step loop and the hint level loop was necessary to enable a finer-gain control and the implementation of adequate instructional strategies at these two different levels of instruction. The solution-step loop basically iterates over the logical steps of an ideal solution and implements strategies that decide which solution step to focus on next. For instance, an instructional strategy guiding the solution-step loop may decide to work with a student on discovering and articulating a particular step of the solution or just skip the step depending on student's knowledge level with respect to the target domain and student's performance on the current problem so far. For example, more advanced students might benefit more from going through the solution at a faster pace in which case "obvious" steps might be

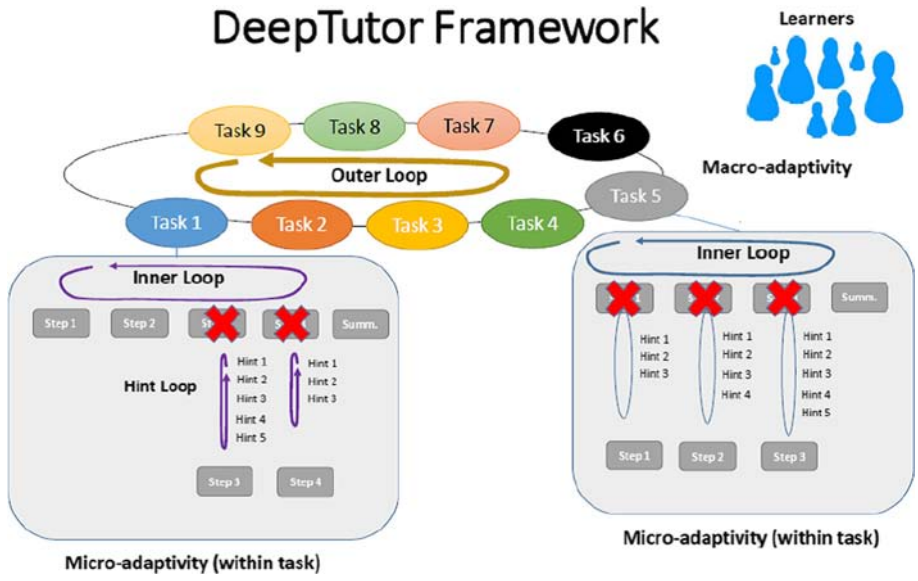

Fig. 2 The 3-loop instructional framework in DeepTutor The 3-loop instructional framework in DeepTutor consists of three major loops: the task loop which loops through the set of tasks in a tutorial session, the solution-step loop which loops through the set of steps in a solution and initiates a hint loop for each missing step in the studnet solution (an X mark on a step in the figure means the student missed to mention that step in their initial solution; the last step is a summarization step in which students are given a summary of the solution), and the hint loop which loops through the set of hints related to a step in a solution

skipped. Once a solution step is set as the next instructional sub-goal, the hint level loop takes care of instructional strategies that scaffold students' self-discovery of that particular solution-step. That is, the role of the hint level loop is to enact strategies that help students construct missing steps in the solution by themselves with minimal help from the system based on constructivist theories of learning. According to these constructivist theories, students construct their own knowledge during learning and the role of the instructor or tutor is to just facilitate this process by offering the right level of support when needed. Accordingly, in ITSs students only get help when needed and in the right doze, i.e. less knowledge students will receive sequences of progressively more informative hints compared to more advanced students who could do well with more vague hints.

We believe that our framework better explains and guides the development of a fully adaptive ITSs. Indeed, having only two loops, the outer loop and the inner loop, is too coarse and obscures important instructional layers that need be addressed explicitly by adopting appropriate instructional strategies as illustrated above for the solution-step and hint level loops. In fact, Rus et al. (2013) suggested there should be even more loops (than the three in our framework) accounted for in a fully independent, comphrehensive, longitudinal education technology that monitors and tutors students over a long period of time spanning many topics and grade levels. According to Rus and colleagues, there should be a loop for each of the following instructional levels: curriculum/standards level, the course level, the lesson level, the activity level, the solution level, and the hint level. Each such loop will have to be guided by different instructional strategies that are appropriate for the corresponding instructional level. For instance, strategies for sequencing instructional tasks across many instructional sessions in a course, which should be informed by principles of interleaving and spacing (Pavlik and Anderson 2008) that have been shown to

promote long-term learning, should be implemented as part of the course level loop in the Rus and colleague's taxonomy of instructional levels. For simplicity and to fairly describe the current state of the DeepTutor system, we only limit our discussion to the three-loop framework mentioned above which addresses the activity (or task) instructional level, solution-step instructional level, and hint instructional level. These three loops are essential in order to understand micro- and macro-adaptivity in DeepTutor which in turn are important to understand the context of our presented here.

The approach

Our approach to predict students' knowledge level in the context of dialogue-based ITSs relies on the fact that each tutorial dialogue between the system and a student has its own characteristics which are strongly influenced by students' background and the nature of instructional tasks. Indeed, students' knowledge level is reflected in the tutorial dialogue between the system and the student, e.g. as the learner becomes more competent the level of help from the ITS should drop. The level of help can be quantified as the number of hints, for instance. Furthermore, the dialogue characteristics are also influenced by the nature of the training tasks. If similar tasks (addressing same concepts in similar or related contexts) are used throughout a whole tutorial session, one might expect that by the time a student reaches the last problems in the session he would master them, thus, requiring less help from the tutor by the end of the session. On the other hand, if the problems are increasingly challenging or simply unrelated to each other then the students would be continuously challenged throughout the whole session; in such a scenario the number of hints a student receives should not drop throughout a session.

We are exploring the relationship between students' prior knowledge and dialogue features in two different setups with two different task selection strategies which allows us to explore the impact of different task selection policies on the dialogue characteristics and therefore on our models for predicting students' prior knowledge. Indeed, we work with data collected from training sessions with two versions of DeepTutor: micro-adaptive-only and fully-adaptive (macro- and micro-adaptive). In the micro-adaptive-only condition, students are working on tasks that were so selected to address typical challenges for all students, i.e. following a one-size-fits-all approach. In this micro-adaptive-only condition, students received scaffolding while working on a task (within-task adaptivity) based on their individual performance on that particular task. For instance, if a student articulated a misconception during the solving of a problem, the system would correct it.

In the macro-adaptive condition, students were assigned to four groups corresponding to four knowledge levels (low knowledge, medium-low knowledge, medium-high knowledge, and high-knowledge) and appropriate instructional tasks were assigned to each group using an Items-Response Theory style analysis (Rus et al. 2014). That is, high-knowledge students received more challenging problems appropriate for their level of expertise while low knowledge students received less challenging problems. The consequence of this more-adaptive task selection policy is reflected in the dialogue characteristics as, for instance, the percentage of hints (explained later) is expected to be similar for both high-knowledge and low-knowledge students as the tasks are similarly challenging relative to the knowledge level of the students. Within a task, the fully-adaptive ITS offered identical micro-adaptivity to the micro-adaptive-only ITS. It should

be noted that in the micro-adaptive-only case, the problems were selected (two each) from the set of problems used for the four knowledge groups in the fully-adaptive condition.

The features of the prediction model

The proposed approach relies on a set of features that was inspired from the previous work mentioned earlier as well as other work such as automated essay scoring (Shermis and Burstein 2003) in which the goal is similar to some extent to ours: infer students' knowledge level or skills based on their language in a written essay. Furthermore, our set of features is grounded in the learning literature as explained next.

The set of dialogue interaction features we employed can be classified into three major categories: time-on-task, generation, and pedagogy. Time-on-task, which reflects how much time students spend on a learning task, correlates positively with learning (Taraban and Rynearson 1998). Time-on-task is measured in several different ways in our case such as total time (in minutes) or normalized total time (we used the longest dialogue as the normalization factor). We computed several additional time-related features such as average time per turn and winsorized versions of the basic time-related features.

Generation features are about the amount of text produced by students. Greater word production has been shown to be related to deeper levels of comprehensions (Chi et al. 2001; VanLehn et al. 2007). We mined from our dialogues many generation-related features such as dialogue length, average turn length, vocabulary size, content word vocabulary size (content words: nouns, verbs, adjectives, and adverbs), and target domain vocabulary size, i.e. a measure of how many words from our target domain, which is Physics, students used.

Lastly, we extracted pedagogy-related features such as how much scaffolding a student received (e.g. number of hints) during the training. Scaffolding is well documented to lead to more learning than lecturing or other, less interactive types of instruction such as reading a textbook (VanLehn et al. 2007). Feedback is an important part of scaffolding and therefore we also extracted features about the type (positive, neutral, negative) and frequency of feedback (Shute 2008).

We extracted raw features as well as normalized versions of the features. In some cases, the normalized versions seem to be both more predictive and more interpretable. For instance, the number of hints could vary a lot from simpler/short problems, where the solution is relatively short and require less scaffolding in general, to more complex problems with longer solutions which require more scaffolding as there are more steps in the solution. A normalized feature such as percentage of hints would allow us to better compare the level of scaffolding in terms of hints across problems of varying complexity or solution length. In our case, we normalized the number of hints by using the maximum number of hints a student may get for a particular problem which happens when the student responds entirely incorrectly to every single hint from the computer tutor. We can infer the largest number of helpful moves, i.e. hints, from our dialogue management component a priori.

Table 1 provides a list of the most important features in each category. For reference, we mined a total of 43 features from 1200 units of dialogue which led to $43 \times 1200 = 51600$ measurements. Our unit of dialogue analysis was a single problem in a training session. Because the training session consisted of 8 problems and we collected 150 sessions from 150 students we ended up with $8 \times 150 = 1200$ dialogue units.

Table 1 Statistics of the dialogue corpus

Category	Features
Time-on-task features	total_time: the time length of the dialogue in minutes
	avg_time_per_turn: the average length of a student turn in minutes
Generation features	dialogue_size: total length of the student dialogue (#words, excl. punctuation)
	avg_dialogue_size_per_turn (#words, no punctuation)
	dialogue_length_div_voc: dialogue_size divided by student's vocabulary size
	#chunks: total number of syntactic constituents or chunks
	#sentences: total number of sentences
	content_vocSize: the vocabulary size of content words
	non_content_vocSize: the vocabulary size of non-content words
	vocSize: total vocabulary size
	%physicsTerms: percentage of physics related terms out of all the words used
	%longWords: percentage of long words out of those used
	%puctuation:percentage of punctuation out of all tokens used
	%articles: percentage of articles such as an or the out of all the words used
	%pronouns: # of non-self-reference pronouns (you, they) out of all words
	%self-references: # of self-reference pronouns (me or we) out of all words
	totalIC: total Information Content of the dialogue
	totalIC_per_word
	positiveness: text positiveness computed based on SentiWordNet
	negativeness: text negativeness
Scaffolding features	#turns: total number of student's turns
	#normalized total number of student turns
	#c_turns: number of student turns classified as contributions (no questions)
	%pos_fb: percentage of turns for which student received positive feedback
	%neg_fb: percentage of turns for which student received negative feedback
	pos_div_pos+neg: positive feedback divided by (positive+negative) feedback
	#shownHints: total number of shown hints
	#shownPrompts: total number of shown prompts, a type of hints
	#shownPumps: total number of shown pumps, a type of hints

The data

As already stated, we conducted our research on log-files collected from an experiment with DeepTutor (Rus et al. 2013). An important aspect of the data we use is the fact that it was collected from system-student interactions outside of the lab. The data was collected during a multi-session, online, after-school experiment (see Fig. 3) in which students interacted with DeepTutor over a period of 5 weeks (one-hour of training per week plus pre- and post-tests). The pre-test and post-test were taken over the web during school hours, under the strict supervision of a teacher. All training sessions were unsupervised as the student chose the type of device (computer, tablet, smartphone), the place (home, library), and time (after-school during week days or during the week-end) to access DeepTutor. This was possible because DeepTutor is a fully-online conversational ITS which can be accessed using a browser from any device with an Internet connection.

Students were encouraged to finish each training session in one sitting. While many did so, some others have finished the training sessions in multiple sittings, spanning several days in a week. We only included in our analyses here students who finished all sessions (all five sessions of the experiment) and did so as instructed, i.e. they finished each session in one sitting. A cohort of 150 students finished everything as instructed out of

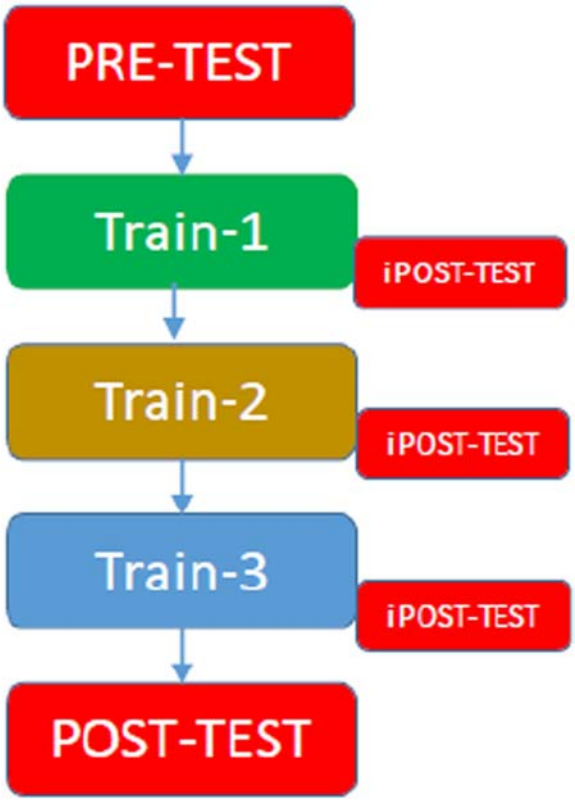

Fig. 3 The Experiment Design The experiment from which we collected data included five sessions: one pre-test session (supervised by teachers), 3 training sessions (fully-online, unsupervised, anytime/anywhere), one post-test session (supervised by teachers). After each training session, student took a short post-test (unsupervised) to measure immediate learning gains

the 365 students who took the pre-test. Each training week about 225 students accessed the system on average. The participants were randomly assigned to one of two conditions mentioned earlier: Micro-Adaptive-only ($n = 70$) and Fully-adaptive ($n = 80$). We only analyzed in this article dialogue corresponding to the first session of training, which focused on the topic of force-and-motion (Newton's first and second law), because it was closest to the pre-test. Table 2 shows summary statistics on this data.

Experiments and results

Our goal was to understand how various characteristics associated with dialogue units corresponding to instructional tasks in a session relate to students' prior knowledge as measured by the pre-test, which is deemed as an accurate estimate of students' prior knowledge level. Our first step towards this goal was to do a feature analysis which is described next.

Table 2 Statistics of the dialogue corpus

Condition	# of complete dialogues	# of dialogue turns	#Sentences
Interactive	80	4587	5102
Adaptive	70	3604	4154
Total	150	8191	9256

Feature analysis

After acquiring all the features for the sub-dialogues corresponding to individual problems our first step was to identify the features whose values best correlate with the pre-test scores. We considered both the entire pre-test (an extended version of Force Concept Inventory; (Hestenes et al. 1992)), but also the portion of the pre-test, which we call pre-test-FM, containing questions directly related to the topic of force-and-motion (FM) training session, i.e. the first training session in our experiment which is our focus as explained earlier. The overall pre-test can be seen as assessing students' overall knowledge with respect to Newtonian Physics, while pre-test-FM is the portion of the pre-test that is directly related to the topic covered during the first week of training, force-and-motion, which basically targeted Newton's first and second laws. Table 3 shows correlations of features with the pre-test scores for the micro-adaptive-only condition.

From Table 3 one can see that with some exceptions for problem 5, the time length (ft1), the total number of sentences (fg7), the number of turns (fs1), and the number of hints (fs11) and prompts shown (fs12) have negative correlations with the pre-test scores, while the average word-length of a turn (fg2) and the percentage of turns receiving positive feedback (fs7) have positive correlations. These findings confirm similar findings from previous studies (VanLehn et al. 2007; Stefanescu et al. 2014). Interestingly enough, the number of sentences students produce seem to be less and less correlated with the pre-test scores as the students advance through the training session.

Correlations for the fully adaptive condition were somehow different due to the fact that there was macro-adaptation involved; that is, the set of tasks were selected differently compared to the micro-adaptive-only condition. We show the best correlated features for the high-knowledge students, for which we had the most data, in Table 4. Two new features, compared to the features in Table 3, seem to be specific for this group in the fully-adaptive condition: fg15 (percent of articles such as an and the) and fg16 (percent of pronouns that are not references such as you and they).

Table 3 Correlations values with pre-test (top) and pre-test-FM (bottom) for the most interesting features on each of the 8 problems in the micro-adaptive-only condition

	1	2	3	4	5	6	7	8
ft1	−0.36	−0.408	−0.141	−0.176	−0.225	−0.136	−0.254	−0.181
	−0.408	−0.333	−0.162	−0.182	−0.256	−0.225	−0.25	−0.219
fg2	0.344	0.262	0.242	0.202	0.213	0.230	0.321	0.236
	0.358	0.221	0.254	0.183	0.157	0.125	0.267	0.216
fg7	−0.423	−0.403	−0.303	−0.295	−0.35	−0.245	−0.283	−0.225
	−0.433	−0.293	−0.268	−0.296	−0.305	−0.258	−0.29	−0.228
fs1	−0.448	−0.444	−0.333	−0.308	−0.34	−0.36	−0.361	−0.276
	−0.473	−0.334	−0.305	−0.295	−0.278	−0.351	−0.331	−0.254
fs7	0.458	0.368	0.193	0.360	0.254	0.208	0.311	0.264
	0.458	0.297	0.122	0.386	0.206	0.168	0.251	0.23
fs11	−0.424	−0.425	−0.215	−0.291	−0.284	−0.326	−0.415	−0.326
	−0.464	-0.314	-0.248	−0.318	−0.223	−0.317	−0.393	−0.29
fs12	−0.404	−0.386	−0.295	−0.352	−0.28	−0.337	−0.194	−0.2
	−0.385	−0.284	−0.225	−0.310	−0.219	−0.304	−0.158	−0.208

Rows correspond to features and columns to problems

Table 4 Correlations values with pre-test (top) and pre-test-FM (bottom) for the most interesting features on each of the 8 problems in the fully-adaptive condition – the high-knowledge group of students

	1	2	3	4	5	6	7	8
fg2	0.144	0.428	0.502	0.652	0.557	0.546	0.367	0.314
	0.193	0.428	0.455	0.646	0.559	0.551	0.224	0.385
fg7	–0.227	–0.604	–0.660	–0.540	–0.357	–0.600	–0.194	–0.254
	–0.482	–0.664	–0.609	–0.545	–0.409	–0.678	0.023	–0.208
fg15	0.181	0.492	0.493	0.443	0.581	0.474	0.588	0.730
	0.002	0.393	0.524	0.350	0.648	0.484	0.614	0.636
fg16	–0.291	–0.329	–0.508	–0.551	–0.528	–0.513	–0.144	–0.754
	–0.446	–0.201	–0.356	–0.332	–0.441	–0.389	–0.139	–0.576
fs1	–0.252	–0.507	–0.668	–0.559	–0.405	–0.603	–0.228	–0.369
	-0.475	–0.629	–0.626	–0.578	–0.438	–0.656	0.029	–0.470
fs11	–0.056	–0.568	–0.569	–0.693	–0.328	–0.530	–0.127	–0.339
	–0.335	–0.63	–0.549	–0.629	–0.303	–0.656	0.078	–0.469
fs12	–0.254	–0.343	–0.602	–0.486	–0.352	–0.409	–0.210	–0.159
	–0.504	–0.534	–0.510	–0.512	–0.087	–0.403	0.013	–0.284

Rows correspond to features and columns to problems

Predicting students knowledge level

To predict students' knowledge levels, we generated regression models from subsets of consecutive problems in a training session in order to understand after how many problems the prediction of students' knowledge level is best. The models were generated not only based on all the available features, but also on subsets of features corresponding to the three major categories of features: Time-on-Task, Generative, and Pedagogy/Scaffolding. All the models were generated using the Backward method in SPSS, so as to be able to find the r value corresponding to the highest adjusted r square value and the lowest degrees of freedom (fewest predictors). It is important to note that in the fully-adaptive condition the models were generated separately for the four groups of students corresponding to the four knowledge levels.

The results in Tables 5 and 6 indicate that after only four problems, the explained variance is comparable to the best case, which is obtained using the first 6 out of the eight tasks. Scaffolding-related features as a group seem to do better, which is very informative and expected in a tutorial context as opposed to non-instructional, general dialogues. Tables 7 and 8 present results for dialogues from the fully-adaptive condition. Similar to the results obtained from the micro-adaptive-only condition, the scaffolding features as

Table 5 r (top) and adjusted r square (bottom) values for cumulative sub-dialogues in the micro-adaptive condition and the pre-test

	1	1–2	1–3	1–4	1–5	1–6	1–7	1–8
All	0.735	.706	.726	.819	.814	.878	.860	.871
	0.451	.426	.458	.600	.589	.693	.669	.678
Time	0.462	.618	.654	.692	.706	.712	.733	.709
	0.193	.358	.406	.436	.465	.481	.507	.483
Gen	0.606	.616	.668	.679	.724	.739	.763	.762
	0.265	.329	.374	.408	.456	.495	.528	.520
Scaf	0.593	.587	.599	.647	.603	.615	.603	.607
	0.289	.310	.306	.343	.302	.308	.311	.317

Table 6 r (top) and adjusted r square (bottom) values for cumulative sub-dialogues in the micro-adaptive-only condition and the pre-test-FM

	1	1–2	1–3	1–4	1–5	1–6	1–7	1–8
All	0.725	.677	.713	.840	.791	.878	.857	.861
	0.449	.405	.438	.611	.552	.693	.667	.681
Time	0.510	.600	.613	.665	.677	.698	.690	.689
	0.260	.335	.360	.404	.422	.452	.448	.454
Gen	0.606	.578	.616	.685	.698	.722	.750	.745
	0.265	.279	.338	.392	.421	.459	.492	.497
Scaf	0.561	.569	.573	.625	.586	.578	.576	.577
	0.259	.278	.273	.322	.249	.279	.287	.288

a group do best. Interestingly, in this case after only two problems the correlation coefficient r is quite high. In a way, this is a validation that the problem selection strategy does a good job at selecting most appropriate problems for each of the four knowledge groups. We know that the problems were appropriately selected because the learning gains for students in the macro-adaptive condition were significantly higher than for students in the micro-adaptive-only condition (the focus of another paper).

Conclusions and future work

We explored in this article models to predict students' prior knowledge based on features characterizing the dialogue-based interaction between a computer-based tutor and a learner. This work was part of our greater goal to move towards non-intrusive assessment methods that would allow learners to focus on the major task, e.g. solving problems or playing a game, and improve their learning experience by eliminating test axieties and tiring effects.

Our results are quite promising with respect to moving towards a world in which learners focus on instruction with no explicit testing. Indeed, our linear regression models based on a number of interaction features yielded in the best cases an $r = 0.949$ and adjusted r-square $= 0.833$. This best result was obtained when developing prediction models using the data from the fully-adaptive ITS. This is expected because in the fully-adaptive case the models were more specialized, i.e. we derived prediction models for each of the four student knowledge levels: low knowledge, medium-low knowledge, medium-high knowledge, and high-knowledge. It should be noted that the best results for the prediction model derived from the micro-adaptive-only ITS data were very good too: $r = 0.878$ and r-square $= 0.693$. Furthemore, scaffolding features seemed to be the

Table 7 r (top) and adjusted r square (bottom) values for cumulative sub-dialogues in the fully-adaptive condition and the pre-test

	1	1–2	1–3	1–4	1–5	1–6	1–7	1–8
All	.820	.910	.949	.930	.922	.935	.936	.924
	.589	.764	.833	.812	.788	.802	.839	.811
Time	.566	.723	.837	.847	.841	.823	.882	.865
	.290	.509	.677	.701	.684	.652	.752	.724
Gen	.782	.862	.904	.906	.896	.850	.909	.884
	.530	.671	.771	.791	.774	.675	.807	.749
Scaf	.709	.662	.838	.769	.754	.752	.782	.804
	.472	.394	.679	.552	.511	.508	.574	.606

Table 8 r (top) and adjusted r square (bottom) values for cumulative sub-dialogues in the fully-adaptive condition and the pre-test-FM

All	.811	.862	.865	.856	.896	.878	.889	.875
	.562	.638	.667	.639	.696	.663	.716	.633
Time	.441	.622	.739	.738	.725	.689	.778	.707
	.158	.349	.510	.510	.489	.434	.560	.452
Gen	.699	.808	.837	.810	.792	.781	.849	.811
	.431	.572	.631	.592	.556	.544	.662	.585
Scaf	.686	.495	.702	.663	.657	.668	.705	.661
	.446	.211	.444	.355	.357	.374	.432	.394

most predictive as a group, as somehow anticipated in a tutorial context, followed by content-generation features.

Our findings have two important implications for the future development of ITSs that would integrate non-intrusive assessment methods such as the ones proposed in this article. *First*, the best models derived from the micro-adaptive-only sessions provide a better estimate of the accuracy ITS developers should expect for predicting learners' prior knowledge level in future ITSs and should be the model to be integrated first in such future ITSs, despite the fact that these models are less accurate, although pretty accurate for that matter, than the more specialized models derived from the fully-adaptive ITS data. The reason is obvious: in order to use the fully-adaptive models, the ITS needs to make a guess or have some a priori measurement of the learners' knowledge, so that it can decide which fully-adaptive model to use for a more precise measurement of learners' knowledge levels based on their performance on the tasks in the tutorial session. However, giving learners a pre-test in order to infer their knowledge level first defies in a way the whole purpose of our intended goal: inferring learners' prior knowledge level from characteristics of the tutor-learner interaction only, without an explicit pre-test. In the case when a learner's knowledge level is known a priori, e.g. from a recent classroom test, and is available as input to the ITS then the ITS could simply trigger the more specialized and more accurate prediction model corresponding to the specific learner's knowledge level without the need to use the micro-adaptive-only prediction model.

Second, the fully-adaptive models' high accuracy can be interpreted as validating the set of selected instructional tasks, i.e. Physics problems in our case, in the tutorial session. Task selection is a critical step in a computer tutor because it has major implications for the effectiveness of the system. If the tasks are too easy, then the learner is bored leading to her disengagement while if the tasks were too difficult the learner would be frustrated and, again, disengaged, to the point that in some cases she might even quit using the tutoring system. Indeed, the tasks should be at the right level of difficulty, not too easy and not too difficult but just right, in order to stimulate the learner and keep her engaged in the learning process throughout the whole tutorial session. That is, the role of the intelligent tutoring system is to keep the learner in the zone of proximal development (Vygotsky 1978) through an appropriate set of tasks with respect to the learner's current knowledge state. In this sense, having components that could monitor the quality of the selected task would thus be very beneficial. It should be noted that because the task selection step is an upstream step in the tutorial process any bad decision regarding task selection would propagate to later, downstream tutoring stages. To illustrate our point, imagine an ITS with a perfect micro-adaptive module which would provide ideal scaffolding to each

learner working on a particular Physics problem. Even if the scaffolding within a task were optimal, learners would not learn much if the Physics problem were way below their knowledge level. Not only that but, as mentioned earlier, the learner would feel bored and in the worst case scenario she might decide to quit using the tutoring system. Our recommendation is that future developers of ITSs should implement both types of models: the micro-adaptive-only models are needed to get a sense of learners' knowledge level without an explicit pre-test while the fully-adaptive models are needed to monitor and validate learners' knowledge level and the quality of the instructional tasks throughout the entire tutorial session.

We plan to further explore the topic of assessing students' prior knowledge from dialogues by investigating affect-related features as well as by using other prediction mechanisms such as classifiers to predicting categorical knowledge levels. Furthermore, we plan to study how similar models can predict post-test scores. We are aware that students' knowledge levels evolve during training, assuming they learn, and therefore there are limitations to our methodology. We do plan to explore in the future ways to infer students' knowledge levels throughout a session, e.g. by having a human expert read the transcripts of a tutoring session.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
This research was supported by the Institute for Education Sciences (IES) under award R305A100875 to Dr. Vasile Rus. All opinions and findings presented here are solely the authors'.

Received: 3 November 2015 Accepted: 22 January 2016

References

BS Bloom, The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educ. Res. **13**, 4–16 (1984)

MTH Chi, SA Siler, H Jeong, Can tutors monitor students' understanding accurately? Cogn. Instr. **22**(3), 363–387 (2004)

MTH Chi, SA Siler, H Jeong, T Yamauchi, RG Hausmann, Learning from human tutoring. Cogn. Sci. **25**(4), 471–533 (2001)

T Corcoran, FA Mosher, A Rogat, Learning progressions in science: An evidence-based approach to reform (CPRE Research Report #RR-63). Consortium for Policy Research in Education. University of Pennsylvania (2009). http://eric.ed.gov/?id=ED506730

K Forbes-Riley, DJ Litman, in *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, Modelling user satisfaction and student learning in a spoken dialogue tutoring system with generic, tutoring, and user affect parameters (Association for Computational Linguistics, New York, New York, 2006), pp. 264–271

AC Graesser, NK Person, JP Magliano, Collaborative dialogue patterns in naturalistic one-to-one tutoring. Appl. Cogn. Psychol. **9**, 495–522 (1995)

D Hestenes, M Wells, G Swackhamer, Force concept inventory. Phys. Teach. **30**(3), 141–158 (1992)

M Lintean, V Rus, R Azevedo, Automatic detection of student mental models during prior knowledge activation in metatutor. Int. J. Artif. Intell. Educ. **21**(3), 169–190 (2012)

L Mohan, J Chen, CW Anderson, Developing a multi-year learning progression for carbon cycling in socio-ecological systems. J. Res. Sci. Teach. **46**(6), 675–698 (2009)

C Moldovan, V Rus, AC Graesser, in *The Proceedings of 22nd Midwest Artificial Intelligence and Cognitive Science Conference*, Automated speech act classification for online chat, (2011), pp. 23–29

P Pavlik, JR Anderson, Using a model to compute the optimal schedule of practice. J. Exp. Psychol. Appl. **14**(2), 101 (2008)

JW Pennebaker, ME Francis, RJ Booth, Linguistic inquiry and word count: Liwc 2001. Mahway: Lawrence Erlbaum Associates. **71**, 2001 (2001)

HL Roediger, JD Karpicke, The power of testing memory: Basic research and implications for educational practice. Perspect. Psychol. Sci. **1**(3), 181–210 (2006)

C Romero, M-I López, J-M Luna, S Ventura, Predicting students' final performance from participation in on-line discussion forums. Comput. Educ. **68**, 458–472 (2013)

R Rus, W Baggett, E Gire, D Franceschetti, M Conley, A Graesser, in *Design Recommendations for Intelligent Tutoring Systems: Learner Modeling*. ed. by R Sottilare, AC Graesser, X Hu, and H Holden, Toward learner models based on Learning Progressions in DeepTutor, vol. 1 (Army Research Laboratory, Orlando, FL, 2013), pp. 183–192

V Rus, S D'Mello, X Hu, A Graesser, Recent advances in conversational intelligent tutoring systems. AI Mag. **34**(3), 42–54 (2013)

V Rus, M Lintean, R Banjade, NB Niraula, D Stefanescu, in *ACL (Conference System Demonstrations)*, Semilar: The semantic similarity toolkit (Citeseer, 2013), pp. 163–168
V Rus, AC Graesser, W Baggett, D Franceschetti, D Stefanescu, N Niraula S Trausan-Matu, K Boyer, M Crosby, K Panou (eds.), *Macro-adaptation in conversational intelligent tutoring matters. Automated response to questions with production rules* (Springer International Publishing, Switzerland, 2014), pp. 242–247
MD Shermis, J Burstein, *Automated Essay Scoring: A Cross-disciplinary Perspective*. (Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, 2003)
VJ Shute, Focus on formative feedback. Rev. Educ. Res. **78**(1), 153–189 (2008)
VJ Shute, M Ventura, *Stealth Assessment: Measuring and Supporting Learning in Video Games*. (MIT Press, Cambridge, 2013)
D Stefanescu, V Rus, AC Graesser, in *Educational Data Mining 2014*, Towards assessing students' prior knowledge from tutorial dialogues (International Educational Data Mining Society, 2014)
R Taraban, K Rynearson, Computer-based comprehension research in a content area. J. Dev. Educ. **21**(3), 10 (1998)
K VanLehn, The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educ. Psychol. **46**(4), 197–221 (2011)
K VanLehn, AC Graesser, GT Jackson, PW Jordan, A Olney, CP Rosé, When are tutorial dialogues more effective than reading? Cogn. Sci. **31**(1), 3–62 (2007)
LS Vygotsky, *Mind in Society: The Development of Higher Psychological Processes*. (Harvard University Press, Cambridge, 1978)
C Williams, S D'Mello, in *Intelligent Tutoring Systems*, Predicting student knowledge level from domain-independent function and content words (Springer, Springer-Verlag Berlin, Heidelberg, 2010), pp. 62–71
BP Woolf, *Building Intelligent Interactive Tutors: Student-centered Strategies for Revolutionizing E-learning*. (Morgan Kaufmann, Elsevier, Burlington, 2008)
J Yoo, J Kim, in *Intelligent Tutoring Systems*, Predicting learner's project performance with dialogue features in online q&a discussions (Springer, 2012), pp. 570–575