Measuring the Power of Learning.™

R & D Connections

No. 25 • October 2015

Conversation-Based Assessment

By G. Tanner Jackson and Diego Zapata-Rivera[1]

Definitions:

Avatar, agent – computer-controlled artificial character

Scaffolding – in education, scaffolding refers to learning support structures designed to help a student understand a concept more fully

Acronyms:

CBA – conversation-based assessment

ITS – intelligent tutoring system

Introduction

Imagine a student working with a tutor for the first time. To better understand what the student knows, the tutor may give problems to solve and then review the student's response. If the response was incomplete or indicative of a misunderstanding, the tutor may ask additional questions and follow up with multiple turns of questions and answers. In some instances, the additional questions may reveal that the student understood the concept deeply but, for whatever reason, had failed to provide a complete answer initially. Such an interactive conversation helps reveal what the student knows and is able to do and areas where more learning is needed. This adaptive method allows the student to fully express his or her knowledge and provides the tutor with more diagnostic information than a non-interactive approach.

These types of open-ended human-to-human conversations can provide great insight and evidence for assessment purposes and may be fairly easy to develop and administer on a small scale. However, scoring human-to-human conversations requires human raters, since current artificial intelligence (AI) technologies cannot yet handle such open-ended conversations. Using human raters is costly and requires significant training and monitoring to maintain acceptable levels of rater agreement. Thus, although these human-to-human conversations provide valuable assessment evidence, they are not easy or financially viable to deploy on a large scale.

Could the same type of interaction take place between a student and a computer? That is the idea behind conversation-based assessment (CBA) systems that involve innovative and interactive tasks framed in engaging and meaningful contexts. Such realistic and meaningful interactions are an example of the new and innovative assessment types that are being developed in response to emerging educational standards and the requirements of a modern and global economy. ETS's Research & Development (R&D) division is currently focusing on collecting and evaluating rich evidence about students' learning in subject areas and also about cross-disciplinary skills, attitudes, and proficiencies in complex areas such as systems thinking, scientific reasoning, argumentation, and English language learning.

Human-to-computer conversations are already used in educational learning games, simulation-based training environments, and intelligent tutoring systems (Millis,

[1] *Editor's note*: The authors are researchers in ETS's Research & Development division. G. Tanner Jackson is a managing research scientist, and Diego Zapata-Rivera is a senior research scientist.

"Human-to-computer conversations are already used in educational learning games, simulation-based training environments, and intelligent tutoring systems."

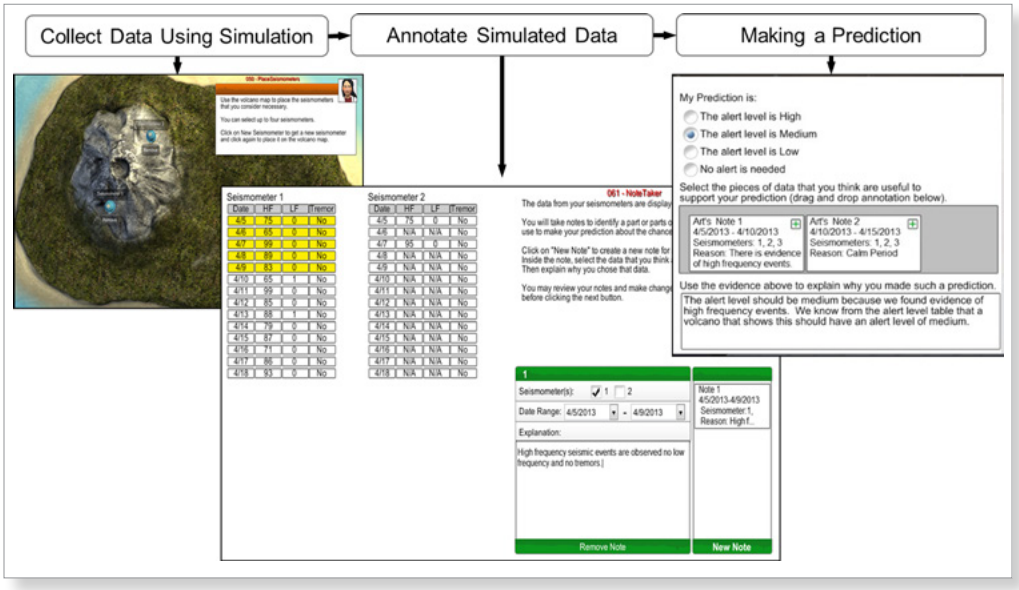
Forsyth, Butler, Wallace, Graesser, & Halpern, 2011; Zapata-Rivera, Jackson, Liu, Bertling, Vezzu, & Katz, 2014). Until recently, the focus has been on learning, but efforts are underway to explore and leverage methods for modeling conversations for assessment purposes. Automated CBA systems are highly scalable because they involve conversing with a computer system (rather than a human) while also providing advantages over traditional assessment systems by integrating the observation and measurement of multiple interacting skills (e.g., cognition, communication, and emotion) within a single, standardized context.

The iterative nature of a conversation allows people to express ideas and understanding in their own words and interact with others in adaptive and appropriate ways. By situating this interaction within an assessment context, CBA allows for a degree of "back-and-forth" reciprocal interactions, effectively leveraging content through conversation and adapting subsequent interactions to target specific information that may be missing from test takers' initial responses. Further, the content of the language generated in conversations can reveal underlying mental models and conceptions, including misconceptions, about complex ideas and processes. In some cases, conversations can more authentically represent the construct targeted for measurement, especially with regard to language or communicative skills.

Origins of Conversation-Based Assessment at ETS

Previous innovative assessment work on scenario-based tasks provided the impetus for pursuing conversation-based assessment at ETS. Specifically, a scenario-based task was developed to assess students' science reasoning skills. This first prototype implemented multiple-choice, constructed-response, and simulation-based items. In the original prototype, test takers were given information about volcanic eruptions, collected sample data on a volcano simulation, and then made a prediction on the likelihood of an eruption based on their data (see Figure 1 for sample screenshots).

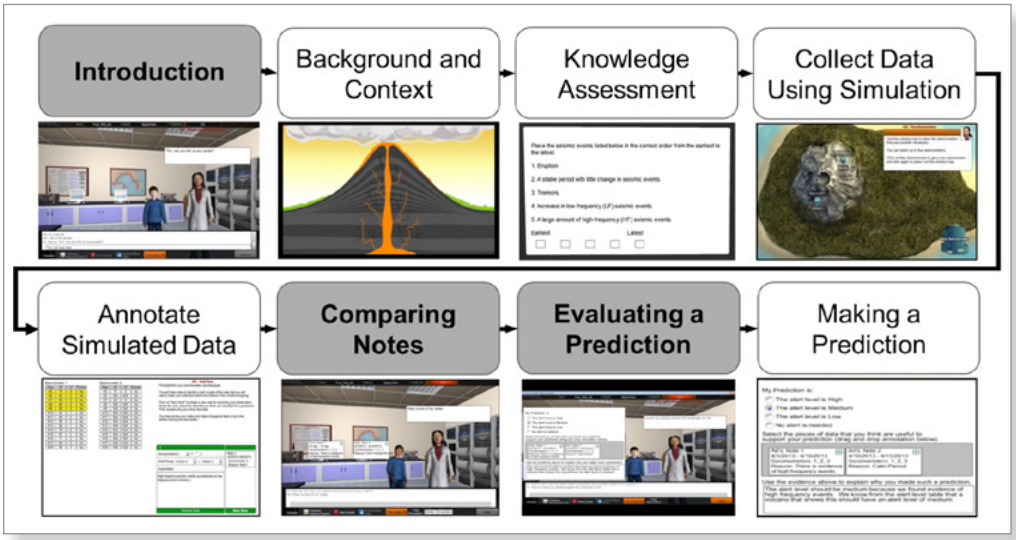
Figure 1.



Users place seismometers (left), annotate data (center), and make a prediction (right).

"Interactions with the original scenario-based task and in-person interviews with the test takers showed that simple conversations with students could reveal a great deal of information about their underlying mental models and understanding behind particular decisions."

Figure 2.



Task flow (shaded boxes represent automated conversation items).

These interactions produced valuable data on test takers' behaviors and decisions, but they did not provide enough information on why decisions were made related to data collection and how the test takers conceptualized information to be used in making the prediction (Zapata-Rivera, Liu, Katz, & Vezzu, 2013). Interactions with the original scenario-based task and in-person interviews with the test takers showed that simple conversations with students could reveal a great deal of information about their underlying mental models and understanding behind particular decisions. Based on this discovery, automated conversations were incorporated into a new version of the same scenario-based prototype (see Figure 2 for original scenario items in white and new conversational items in grey).

The updated prototype has been implemented and tested with a variety of target users (e.g., middle school students). These conversations provide qualitatively different evidence than what is being gathered from the more conventional format items, and work is currently being conducted that explores how to combine this conversation-based evidence using coherent scoring models that represent a more complete picture of what students know and can do.

Foundational Work

ETS's CBA work leverages more than a decade of research on intelligent tutoring systems (ITSs) that involve conversations between a human and virtual characters (Biswas, Schwartz, Bransford, & Teachable Agent Group at Vanderbilt, 2001; Chan & Baskin, 1990; Graesser, Person, Harter, & Tutoring Research Group, 2001; Johnson, Rickel, & Lester, 2000; Yang & Zapata-Rivera, 2010).

Some of these ITSs take the form of conversations involving multiple participants consisting of one human and two virtual characters (e.g., a virtual mentor and a virtual peer). Butler, Forsyth, Halpern, Graesser, & Millis (2011) describe conversation-based tasks as a way to create learning environments that simulate particular pedagogical strategies or social interactions.

"Conversations are a natural fit for assessment because they must accurately assess users' knowledge and skills in order to adaptively respond to their input."

Recent research applies the conversational approach to assessment (e.g., Operation ARIES! in Millis et al., 2011; also see Volcano Scenario in Zapata-Rivera et al., 2014). Conversation-based tutoring systems are a natural fit for assessment because they already include an assessment component; tutoring systems must accurately assess users' knowledge and skills in order to adaptively respond to their input. These systems include "conversational patterns" — alternative paths or threads that are taken based on the participant's input within a dialogue — that are designed to elicit evidence about what the test taker knows or can do. Such evidence may be different from what would result from more traditional measures, in part because students may have multiple opportunities and scaffolds to help them produce a complete answer.

Creating Rich Situations for Assessments

Creating assessments based on conversations presents a serious challenge. A CBA task must, on the one hand, encourage test takers to use a sufficiently rich language that can provide evidence of their knowledge and skills. On the other hand, for automated analysis of natural language conversations, the scenarios must be constrained enough so that they can be scored reliably and used to make valid interpretations of the results. These conversations can take a variety of forms, and specific implementations vary based on the purpose and goal for each task.

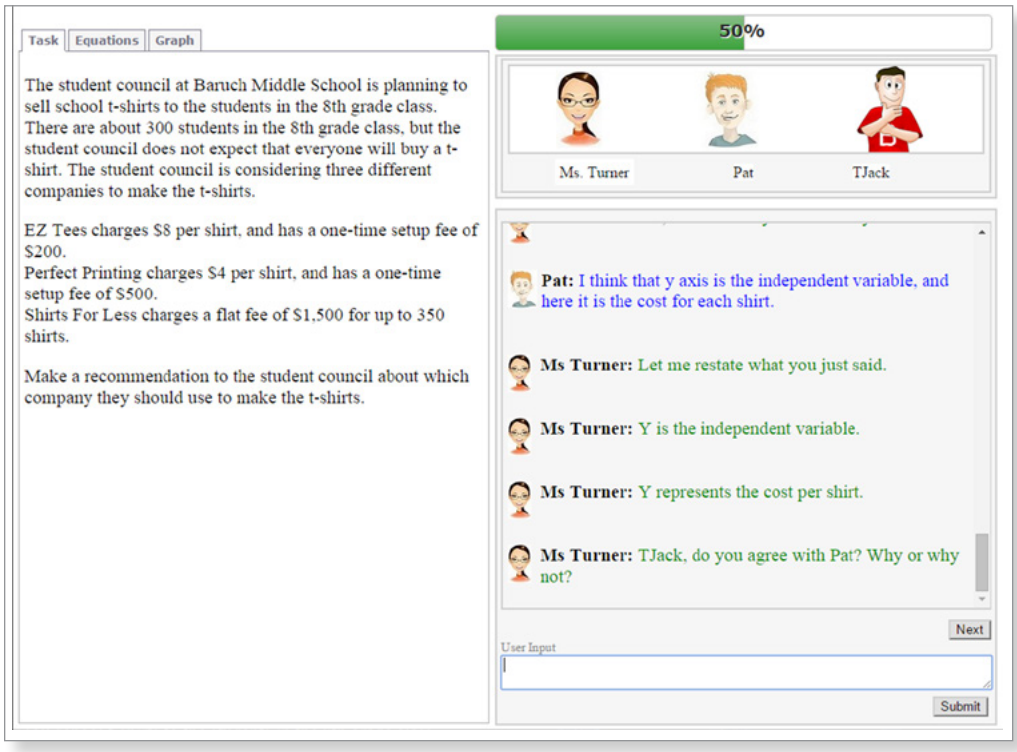
The examples displayed in Figures 3–5 illustrate different approaches to using conversation as the evidentiary basis for making claims about test takers' knowledge and skills. Each of these examples incorporates computer-controlled artificial characters (i.e., agents or avatars). This approach facilitates standardization and improves reliability by controlling the range of potential construct-irrelevant factors, which could appear if the conversation goes into areas that are not relevant to the test. However, the flip side of this is that if a student provides a novel (i.e., unanticipated) but correct response, the system may not be able to reliably assess it.

The current CBA developments at ETS utilize the conversation-based approach and leverage the benefits of interacting with multiple agents (Figures 3–5). These conversations are carefully designed to provide opportunities for individuals to give evidence of their knowledge and skills, as well as to provide scaffolding for the learning process and give the test taker valuable feedback. In addition, care is taken to ask questions in such a way that current AI technologies can process student responses. These tasks embody a range of implementations and uses that may be conceptualized as a series of constructed-response items (Figure 3), represented as a simple chat room environment (Figure 4), or even integrated into immersive environments to elicit evidence of target skills (Figure 5). For example, in Figure 5, interactions occur between the human test taker and the two virtual student agents as they try to work through math problems collaboratively. In this figure, the human user has answered the question incorrectly and the conversation system selects a path such that each of the virtual characters also chooses a different answer so that everyone has a different opinion. However, before providing direct feedback and asking the participant to answer the question again, this system engages the user in a conversation about the original math problem as a way to detect if there are language difficulties in addition to difficulties with the math content itself. By doing so, this diagnostic assessment assesses and discriminates language and math competencies.

Figure 3.

Human-to-computer automated conversation. (Mandy and Carol are computer avatars engaging with Jackson, a student being assessed on whether he can identify flaws in Mandy's argument.)

Figure 4.



Human-to-computer chat room environment. (Ms. Turner and Pat are avatars interacting with TJack to assess his understanding of elements in this math problem.)

"Some of the critical issues currently being investigated include aspects of CBA validity, reliability, fairness, feasibility, scalability, and generalizability."

Figure 5.

Human-to-computer immersive environment. (Lucas, left, and Sarah, right, are avatars interacting with the human user to assess his or her understanding of a math problem.)

Current and Future Work

Our current research efforts focus on evaluating critical aspects of automated CBA and how the approach could be scaled up for use by a wider audience (K–12 to workforce) and implemented within potential future products. Some of the critical issues currently being investigated include aspects of CBA validity, reliability, fairness, feasibility, scalability, and generalizability. The capability to develop CBAs includes the ability to design, implement, and score naturalistic, communicative, interactive tasks that simulate learning processes and social interactions. Each of these components must be examined, evaluated, and made more efficient in order to scale up the overall CBA approach. To identify and improve aspects for scaling development, recent work has focused on six CBA prototypes, including science reasoning (Figure 2), general argumentation (Figure 3), linear functions and mathematical argumentation (Figure 4), English communication (for English language learners [ELLs]), diagnostic assessments for ELLs in English and mathematics (proportional reasoning; Figure 5), and collaborative problem solving. Across these prototypes, we are investigating several key research issues, primarily around the validity and scalability of this approach. Our research questions include the following:

- *Fairness and Bias*: Do these tasks introduce bias for ELLs, test takers with disabilities, or other subgroups?
- *Generalizability*: Do scores generalize across tasks/forms?

"CBAs have great promise to extend existing assessment approaches and provide new kinds of evidence. They offer a dynamic and iterative process that generates evidence to reveal test takers' abilities to recall, understand, apply, explain, generalize, and communicate important knowledge and skills."

- *Scoring and Scaling*: Can the generated data be efficiently scored and reliably scaled?
- *Construct Representation*: Do the scores reflect contemporary models of our constructs?
- *Impact*: Do these new assessments have their intended positive impact?

We will also research how this approach can generate evidence that could be used as part of summative high-stakes assessments, as well as to provide actionable information to students and teachers about learning within a formative assessment process.

Conclusion

CBAs have great promise to extend existing assessment approaches and provide new kinds of evidence. They offer a dynamic and iterative process that generates evidence to reveal test takers' abilities to recall, understand, apply, explain, generalize, and communicate important knowledge and skills. They have the potential to positively impact teaching and learning through their alignment with complex cognitive and sociocultural models of learning and pedagogy. Advances in technology, educational theories, and psychometrics can be combined in CBAs to help achieve this goal. CBAs can also be used for formative assessment to help test takers understand what they know well and areas where they need to study more or need additional instruction. Such assessments can also be delivered on computers as self-guided assessments, thus reducing teacher time for administration and scoring. If we are successful in our efforts, future assessments that leverage conversational approaches may embody, measure, and support effective learning better than ever before and will provide stakeholders with critical information to enhance deeper learning.

References

Biswas, G., Schwartz, D., Bransford, J., & Teachable Agent Group at Vanderbilt. (2001). Technology support for complex problem solving: From SAD environments to AI. In K. D. Forbus & P. J. Feltovich (Eds.), *Smart machines in education: The coming revolution in educational technology* (pp. 71–97). Menlo Park, CA: AAAI/MIT Press.

Butler, H. A., Forsyth, C., Halpern, D. F., Graesser, A. C., & Millis, K. (2011). Secret agents, alien spies, and a quest to save the world: Operation ARIES! engages students in scientific reasoning and critical thinking. In R. L. Miller, R. F. Rycek, E. Amsel, B. Kowalski, B. Beins, K. Keith, & B. Peden (Eds.), *Promoting student engagement: Vol. 1. Programs, techniques and opportunities* (pp. 286–291). Syracuse, NY: Society for the Teaching of Psychology.

Chan, T. W., & Baskin, A. B. (1990). Learning companion systems. In C. Frasson & G. Gauthier (Eds.), *Intelligent tutoring systems: At the crossroads of artificial intelligence and education* (pp. 6–33). Norwood, NY: Ablex.

Graesser, A. C., Person, N., Harter, D., & Tutoring Research Group. (2001). Teaching tactics and dialog in AutoTutor. *International Journal of Artificial Intelligence in Education*, *12*, 257–279.

Johnson, W. L., Rickel, J. W., & Lester, J. C. (2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education*, *11*, 47–78.

Millis, K., Forsyth, C., Butler, H., Wallace, P., Graesser, A. C., & Halpern, D. (2011). Operation ARIES! A serious game for teaching scientific inquiry. In M. Ma, A. Oikonomou, & J. Lakhmi (Eds.), *Serious games and edutainment applications* (pp.169–196). London, UK: Springer-Verlag.

Yang, H. C., & Zapata-Rivera, D. (2010). Interlanguage pragmatics with a pedagogical agent: The request game. *Computer Assisted Language Learning*, *23*, 395–412.

Zapata-Rivera, D., Jackson, T., Liu, L., Bertling, M., Vezzu, M., & Katz, I. R. (2014). Assessing science inquiry skills using trialogues. In S. Trausan-Matu, K. Boyer, M. Crosby, & K. Panourgia (Eds.), *Intelligent tutoring systems* (Vol. 8474, pp. 625–626). Cham, Switzerland: Springer International Publishing.

Zapata-Rivera, D., Liu, L., Katz, I. R., & Vezzu, M. (2013). Exploring the use of game elements in the development of innovative assessment tasks for science. *Cognitive Technology*, *18*, 43–50.

R&D Connections is published by

ETS Research & Development
Educational Testing Service
Rosedale Road, 19-T
Princeton, NJ 08541-0001
email: **RDWeb@ets.org**

Editor: Hans Sandberg
Copy Editor: Eileen Kerrigan
Layout Design: Sally Acquaviva

Visit ETS Research & Development on the web at **www.ets.org/research**

Follow ETS Research on Twitter® **(@ETSresearch)**

Copyright © 2015 by Educational Testing Service. All rights reserved. ETS and the ETS logo are registered trademarks of Educational Testing Service (ETS). MEASURING THE POWER OF LEARNING is a trademark of ETS. All other trademarks are property of their respective owners. 32208