Measuring the Power of Learning.®

Research Memorandum
ETS RM–18-13

Identifying and Addressing Unexpected Responses in Conversation-Based Assessments

Diego Zapata-Rivera

Blair Lehman

Jesse R. Sparks

Han-Hui Por

Kofi James

December 2018

ETS Research Memorandum Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Memorandum series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Memorandum series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Identifying and Addressing Unexpected Responses in Conversation-Based Assessments

Diego Zapata-Rivera, Blair Lehman, Jesse R. Sparks, Han-Hui Por, and Kofi James
Educational Testing Service, Princeton, New Jersey

December 2018

Corresponding author: D. Zapata-Rivera, E-mail: DZapata@ets.org

Suggested citation: Zapata-Rivera, D., Lehman, B., Sparks, J. R., Por, H.-H., & James, K. (2018). *Identifying and addressing unexpected responses in conversation-based assessments* (Research Memorandum No. RM-18-13). Princeton, NJ: Educational Testing Service.

Action Editor: Keelan Evanini

Reviewers: Malcom Bauer and Vikram Ramanarayanan

Abstract

Adaptive computer systems show great potential in education. Successful adoption of these systems, however, requires researchers to identify and address situations that negatively influence students' learning and assessment experiences. We have designed conversation-based assessments (CBAs) to assess constructs such as science inquiry, English language, and collaborative problem-solving skills. These computer-based conversations have provided useful evidence of students' skills. However, in some cases, students provide unexpected responses that are indicative of a state such as disengagement or negative experience. In this research memorandum, we explore an approach for identifying and addressing these unexpected cases. The proposed approach involves identifying cases of unexpected responses, analyzing cases and defining categories, designing possible solutions, and validating case categories and possible solutions with expert teachers. Results of applying this approach in the context of CBAs suggest that the envisioned solutions could be used to successfully deal with the cases identified. This work will inform the creation of detectors to be used across CBAs to successfully identify and appropriately address unexpected responses.

Key words: Unexpected responses, conversation-based assessments, teacher validation

Acknowledgments

We would like to thank Jonathan Steinberg for his contributions to this work, and the teachers who participated in our study.

Dialogue-based systems have been used successfully in learning contexts. Recent research has shown promise in assessment as well (Graesser, Cai, Morgan, & Wang, 2017; Zapata-Rivera, Jackson, & Katz, 2015). Conversation-based assessments (CBAs) are thought to be an effective method of assessment because the interactional style allows students the opportunity to provide answers in their own words, which (a) allows for a more accurate assessment of student knowledge and (b) creates a more engaging experience for students. However, it is not always the case that CBAs achieve these two results. For example, in some cases, students may provide unexpected responses that are not handled by the system or may be indicative of a state such as disengagement or negative experience.

Approaches to identifying and addressing unexpected responses have been explored in areas such as dialogue-based systems that involve natural-language human–agent communication (Olney et al., 2003; Rus, D'Mello, Hu, & Graesser, 2013). Jordan, Litman, Lipschultz, and Drummond (2009) explored the detection and repair of misunderstandings in dialogue systems. They associated these misunderstandings with problems understanding implicit intentions, not utilizing all of the information available to understand the intended interpretation, and problems connecting the system's request with previous dialogue. Johnson, Ashish, Bodnar, and Sagae (2010) showed that many problems found in real settings are not found in the lab.

These unexpected responses can stem from a number of causes. For example, a student response may not be evaluated correctly, which causes the student to receive inappropriate feedback; the student may dislike something about the task (e.g., topic, interaction style, agent appearance); or the student may be affected by other external factors (e.g., the student does not want to participate in the activity). In the context of dialogue systems designed for assessment, such negative responses represent a form of construct-irrelevant variance that can affect the validity of the evidence gathered from those interactions about students' knowledge, skills, and abilities (Messick, 1994) and are therefore critical to identify and address.

The main goal of this line of research is to learn from existing data to develop strategies that allow us to identify (and possibly prevent) situations that negatively contribute to the assessment goals of the system. CBAs are intended to provide students with multiple opportunities to demonstrate their knowledge and/or skills using simulated, natural-language conversations with virtual agents to elicit explanations about decisions that students make in scenario-based tasks, simulations, or games (Zapata-Rivera et al., 2014).

Even though CBAs have been used to collect evidence of students' knowledge and skills in various domains, such as science inquiry, English language, and collaborative problem solving, in some cases, students do not experience the system in the way it was originally designed to be experienced, making it difficult to use the evidence collected from those students for assessment purposes. In this research memorandum, we describe an approach to identifying and addressing these cases. The proposed approach involves the following activities: (a) identifying cases of unexpected responses, (b) analyzing cases and defining categories, (c) designing possible solutions, and (d) validating case categories and possible solutions with expert teachers. The approach has been applied using existing data from prior CBA research. We discuss lessons learned and elaborate on future work in this area.

Identifying and Addressing Unexpected Responses

This section describes the proposed approach and the results obtained when applying it to data collected using CBAs.

Identifying Cases

The goal of this activity is to identify potential cases of unexpected responses. This can be done by looking at sequences of responses between the student and one or more virtual characters that ended up with negative or unexpected statements (e.g., "I do not want to do this," "Stop asking," "I already answered") or by looking at other available data (e.g., students' self-reported emotions, usability data, and student responses to other tasks).

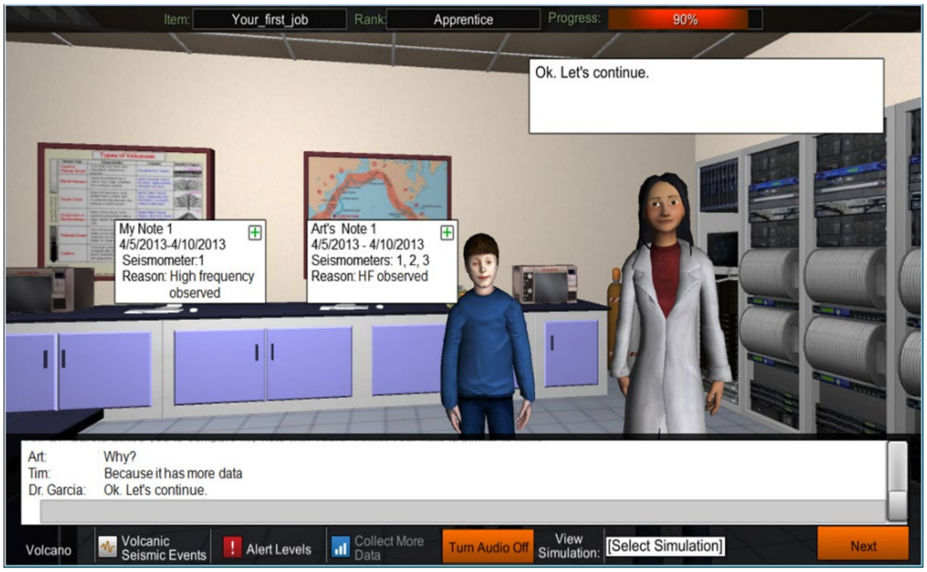
In the current project, we looked at data collected using five CBAs: the science inquiry Volcano (n = 200) and Weather scenarios (n = 500), a math CBA scenario (n = 400), an English-language scenario (n = 80), and an English–math scenario (n = 40). Data for this analysis were based on text responses to written conversations and related context data. Spoken conversations were not included in this analysis. These conversations make use of the ASATA dialogue engine (Nye, Graesser, & Hu, 2014). More information about the authoring process of CBAs can be found in Zapata-Rivera et al. (2015).

A subset of 72 unexpected response cases from 59 students were selected for this study. Team members tracked down cases by looking at sequences that ended up with negative or unexpected statements using conversation logs, students' self-reported emotions, usability data, and student responses to other tasks. The members of the team had expertise in areas such as cognitive science, educational psychology, dialogue systems, and computer science. Cases were tagged with possible causes and the source of data used to identify each case. These cases were tagged with possible causes and the data source.

Figure 1 shows a fragment of a selected case. In this case, Art is an artificial peer agent working with the student and Dr. Garcia is an artificial scientist agent (serving as the teacher/authority figure) within a science inquiry task. The student does not seem engaged in the task, as seen in the student's short, general responses, which do not seem to answer the question posed by the peer. Figure 2 shows a screenshot of the volcano scenario. The student discusses with Art and Dr. Garcia the quality of observation notes.

Art:	[Student], please look carefully at and compare these two notes. I think we should keep your note for making predictions later. Do you agree with me? Why or why not?
Student:	Yep.
Art:	Please tell me more about which note you want to keep and why.
Student:	Sure.
Dr. Garcia:	Ok. Let's continue.

Figure 1. Sample case.



Item: Your_first_job Rank: Apprentice Progress: 90%

Ok. Let's continue.

My Note 1
4/5/2013-4/10/2013
Seismometer:1
Reason: High frequency observed

Art's Note 1
4/5/2013 - 4/10/2013
Seismometers: 1, 2, 3
Reason: HF observed

Art:	Why?
Tim:	Because it has more data
Dr. Garcia:	Ok. Let's continue.

View Simulation: [Select Simulation]

Figure 2. A screenshot of the volcano conversation-based assessment scenario. The student (Tim) explains why he wants to keep Art's note for making predictions about a possible volcanic eruption (because it has more data, meaning data from three seismometers).

Analyzing Cases and Defining Categories

Using available information about each case and the tags described earlier (i.e., possible causes and data source), the research team met to discuss the different types of cases observed and, from this discussion, generated a list of 11 categories (see Table 1). These categories were applied to exemplar cases and agreed upon by the team.

Table 1. Unexpected Response Categories

Category	Description
Confused	The response makes little sense or lacks an explanation, because the student may have insufficient knowledge (e.g., "I do not understand").
Frustrated	The student seems annoyed by his or her interaction with the characters (e.g., "I already said that").
Repetitive	The student repeats the same answers to slightly different questions across multiple turns of the conversation.
Unmotivated	The student does not care to answer: He or she provides short answers due to lack of motivation (e.g., "no," "idk," "sure," "yep") or no answers are provided (including gibberish responses).
Irrelevant	The student is not answering the right question or the answer is off topic (e.g., "Because I do").
Asking for help	The student is asking the characters for addition information (e.g., to repeat a question or show additional materials).
Gaming the system	The student tries to test the capabilities of the system (e.g., by entering profanity).
Attempting to communicate	The student does not answer the question. The student refers to prior dialogue or task content (e.g., "I answered a similar question before").
Using different languages	The student answers the question in a language other than English.
Giving up	The student expresses a wish to quit the task (e.g., "Make it stop please").
Other	This applies when the case does not fit any of the preceding categories.

Designing Possible Solutions

Possible solutions for each of the identified categories were devised. These solutions involve strategies such as changing the type of question, changing the source of the question (e.g., make Dr. Garcia ask the question instead of the peer), intervening by asking the student to please answer the question, reacting to obscene language, and confirming with the student that he or she wants to quit and giving the student an opportunity to do so. Figures 3 and 4 show examples of possible solutions to two particular cases: unmotivated (the case presented in Figure

1) and giving up (a student who does not wish to elaborate on when to choose one printing company over another).

Art:	[Student], please look carefully at and compare these two notes. I think we should keep your note for making predictions later. Do you agree with me? Why or why not?
Student:	Yep.
***Dr. Garcia**:*	*Please answer Art's question.*
Student:	[Response]
Dr. Garcia:	Ok. Let's continue.

Figure 3. Sample solution: Unmotivated.

Virtual Teacher:	Let's keep going.
Virtual Peer:	When is it best to go with Shirts for Less?
Student:	Can you stop please?
Virtual Teacher*:*	*I understand you want to stop. Is that right?*
Student:	[Response]
Virtual Teacher*:*	If Yes, [END] else [Skip to next conversation]

Figure 4. Sample solution: Giving up.

Validating Case Categories and Possible Solutions With Expert Teachers

To validate the proposed approach, a group of three expert teachers were asked to review each case and provide feedback on the categories and potential solutions based on their expertise teaching students in the target population.

Participants. Three expert teachers with a background in teaching middle and high school science, math, and English language arts, respectively, participated in this study. Their teaching experience ranged from 10 to more than 29 years as classroom teachers. Two teachers were enrolled in doctoral programs at the time of the study, and the third held a bachelor's degree. All three teachers had several years of experience working as a rater for constructed-response assessments. Two teachers reviewed each case.

Method. Teachers received training in understanding the CBA approach and the data collection materials developed for this study, which included a description of the CBA task, conversation transcripts with highlighted unexpected responses, and annotations to help teachers understand the context of the conversation. Possible solutions for each unexpected case were also provided to elicit teacher feedback on these solutions.

After reviewing each conversation case between the student and the virtual agents, teachers chose up to three rank-ordered categories (from the list in Table 1) for each highlighted unexpected case within the conversation. Then, they assessed the appropriateness level of the proposed solution using a 4-point Likert-type scale ranging from 1 (*very inappropriate*) to 4 (*very appropriate*) and an open-ended item to explain their selection ("Why?"). Finally, teachers were asked to respond to the questions "If you were the teacher, how would you handle each situation?" and "Is there another point in the conversation where you would intervene? If so, please give the line number within the conversation." A total of 144 responses were collected.

Results. Data collected from expert teachers showed that in 85% of cases, teachers considered the proposed way to handle unexpected cases to be appropriate or very appropriate. Teacher classification of each case using the categories in Table 1 was consistent with our assigned categories. The original category assigned by the research team was among the categories selected by both teachers in 43% of the cases, among the categories selected by only one of the teachers in 46% of the cases, and was not included in 11% of the cases. In 89% of the cases, at least one expert selected the category selected by the research team.

In 37 of the 144 cases, experts chose to use *other*. In most cases, the experts elaborated on the classifications they had already assigned: Examples are frustrated, asking for help, attempting to communicate, and confused. In the remaining cases, the experts used the *other* classification to explain or assign possible causes to the unexpected responses, such as indicating that the student was already answering to the best of his or her ability or "The student feels like the system is not listening to his or her responses." In summary, the *other* responses helped to define the existing classifications in Table 1, and no new classifications were identified.

In 74% of the 72 paired responses, teachers agreed with each other on at least one classification. The most common classifications used by the teachers were unmotivated, frustration, and confused. In 1 case, the two experts agreed on the three categories; in 14 cases, they agreed on two categories; in 31 cases, they agreed on one category; and there were 26 cases of no agreement (10 of which were due to missing data where the expert did not choose any of the categories). This finding indicates that classifying the responses was difficult even for human experts.

Analysis of interrater agreement of the presence or absence of individual categories of expert ratings showed low to moderate kappa values (values between −.16 and .45), with frustration (.45), unmotivated (.31), gaming the system (.17), giving up (.13), and other (.23) having the higher kappa values.

Teachers also provided suggestions for improving the conversations. These suggestions included changes to the wording of interactions/responses (e.g., Dr. Garcia: “We have a zero-tolerance policy toward obscene language”); providing additional guidance to capture students’ attention (e.g., Virtual Teacher: “Let’s look at the equation”); focusing on the task when the student seems to have a problem with one of the characters (e.g., Student: “I do not like Art”; Dr. Garcia: “Let’s please stay on topic”); intervening earlier to prevent downstream issues; making the characters react to occurrences of obscene language; using shorter sentences to facilitate understanding; providing verification feedback; and providing opportunities to verify that the student has a sufficient understanding of the task before asking him or her to answer the question again.

The teachers also cautioned against redirecting questions to the peer character or allowing the peer character to intervene before students had a chance to demonstrate their understanding or before the teacher had a chance to provide adequate feedback to the students’ previous responses. In some cases, teachers suggested the use of open-ended prompts to understand the source of students’ confusion (e.g., “What do you need me to explain?”). Although these open-ended statements are intuitive in a real classroom, they might prove to be challenging to implement because of the limited number of possibilities that can be handled by the system.

In 45.1% of the cases, the teachers would not intervene in addition to the proposed solution. In the remaining cases, experts typically suggested intervening earlier to prevent an escalation of the issues or to provide additional feedback to aid the students’ understanding.

Discussion

The categories of unexpected responses show that in some cases, the dialogue system within CBA tasks did not react appropriately to situations that could result in negative experiences for students. These results can inform the development of new versions of the system. In fact, some causes of confusion have already been identified and fixed, for example, reacting appropriately to statements classified as “other” by the system (e.g., instead of asking “How is this related to our conversation?”—which was inappropriate in cases where students included partially relevant information—rephrasing the request to elicit more information, such as “Please tell me more”). However, other causes of negative reactions may deal with students’ unreasonable expectations about the capabilities of the system (e.g., when students’ expectations are either too high or too low) or the appearance and/or voices of the characters. These issues can be handled by establishing a common ground with the students or acknowledging limitations of the system (e.g., “Sometimes the system may not understand you; in this case, try to state your response in a different way”).

Other issues could be handled by implementing detectors that react to particular situations and prevent potential negative situations from escalating (Baker, Corbett, Roll, & Koedinger, 2008). In some cases, simple detectors suffice. For example, sentiment analysis of written conversations between two humans interacting with a virtual character has shown that sentiment detectors could be used to deal with cases where humans lack trust in the agent’s capabilities or do not like the agent’s appearance (Hao et al., 2018). However, appropriate system reactions for more challenging cases should consider not only the goals of the system but also information about the task, the student interaction history, and the state of the student model.

The findings of this work are limited to the CBA data collected using the CBA prototypes described herein. It is possible that some of the categories analyzed here are not present in other CBAs or other types of interactive tasks.

Related work includes approaches that make use of expert input to annotate cases and improve the performance of models that predict student disengagement (Paquette, de Carvalho, & Baker, 2014). Also, these types of approaches can be used to provide information that helps identify and explain other forms of off-task behavior (Baker, Corbett, Koedinger, & Wagner, 2004; Rowe, McQuiggan, Robison, & Lester, 2009).

Conclusions and Future Work

Results of applying this approach provided valuable information to inform our next steps. These findings can also inform the development of a new generation of assessment systems that can listen to the user and respond appropriately.

Future work will explore the use of detectors that can successfully identify and appropriately address unexpected responses. In addition, we will use the feedback from teachers to refine the unexpected response categories, identify potential solutions to particular cases, and conduct studies to evaluate these revisions with students.

References

Baker, R. S., Corbett, A., Koedinger, K., & Wagner, A. (2004). Off-task behavior in the cognitive tutor classroom: When students "game the system." In *Proceedings of the ACM Conference on Human Factors in Computing Systems* (pp. 383–390). New York, NY: ACM.

Baker, R. S., Corbett, A., Roll, I., & Koedinger, K. (2008). Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction, 18*, 287–314. https://doi.org/10.1007/s11257-007-9045-6

Graesser, A. C., Cai, Z., Morgan, B., & Wang, L. (2017). Assessment with computer agents that engage in conversational dialogues and trialogues with learners. *Computers in Human Behavior, 76*, 607–616. https://doi.org/10.1016/j.chb.2017.03.041

Hao, J., Zapata-Rivera, D., Graesser, A. C., Cai, Z., Hu, X., & Goldberg, B. (2018). Towards an intelligent tutor for teamwork: Responding to human sentiments. In *Design recommendations for intelligent tutoring systems: Vol. 6. Team learning and taskwork.* (pp. 151–160). Adelphi, MD: U.S. Army Research Laboratory.

Johnson, L. W., Ashish, N., Bodnar, S., & Sagae, A. (2010). Expecting the unexpected: Warehousing and analyzing data from ITS field use. In *ITS 2010* (pp. 352–354). New York, NY: Springer.

Jordan, P., Litman, D., Lipschultz, M., & Drummond, J. (2009). Evidence of misunderstandings in tutorial dialogue and their impact on learning. In *Artificial intelligence in education* (pp. 125–132). Amsterdam, Netherlands: IOS Press.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessment. *Educational Researcher, 23*(2), 13–23. https://doi.org/10.3102/0013189X023002013

Nye, B. D., Graesser, A. C., & Hu, X. (2014). AutoTutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education, 24*, 427–469. https://doi.org/10.1007/s40593-014-0029-5

Olney, A., Louwerse, M., Mathews, E., Marineau, J., Hite-Mitchell, H., & Graesser, A. (2003). Utterance classification in AutoTutor. In *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing* (Vol. 2, pp. 1–8). https://doi.org/10.3115/1118894.1118895

Paquette, L., de Carvalho, A. M. J. A., & Baker, R. S. (2014, July). *Towards understanding expert coding of student disengagement in online learning*. Paper presented at the 36th Annual Cognitive Science Conference, Quebec City, Canada.

Rowe, J. P., McQuiggan, S. W., Robison, J. L., & Lester, J. C. (2009). Off-task behavior in narrative-centered learning environments. In *Artificial intelligence in education* (pp. 99–106). Amsterdam, Netherlands: IOS Press.

Rus, V., D'Mello, S., Hu, X., & Graesser, A. C. (2013). Recent advances in intelligent systems with conversational dialogue. *AI Magazine, 34*, 42–54. https://doi.org/10.1609/aimag.v34i3.2485

Zapata-Rivera, D., Jackson, T., & Katz, I. R. (2015). Authoring conversation-based assessment scenarios. In R. A. Sottilare, A. C. Graesser, X. Hu, & K. Brawner (Eds.), *Design recommendations for intelligent tutoring systems: Vol. 3. Authoring tools and expert modeling techniques* (pp. 169–178). Adelphi, MD: U.S. Army Research Laboratory.

Zapata-Rivera, D., Jackson, T., Liu, L., Bertling, M., Vezzu, M., & Katz, I. R. (2014). Science inquiry skills using trialogues. In *12th international conference on Intelligent Tutoring Systems* (pp. 625–626). Honolulu, HI: Springer International Publishing.