Research Report

Examining the Accuracy of a Conversation-Based Assessment in Interpreting English Learners' Written Responses

ETS RR–21-03

Alexis A. Lopez
Danielle Guzman-Orth
Diego Zapata-Rivera
Carolyn M. Forsyth
Christine Luce

December 2021

ETS Research Report 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 Report 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 Report 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.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Examining the Accuracy of a Conversation-Based Assessment in Interpreting English Learners' Written Responses

Alexis A. Lopez, Danielle Guzman-Orth, Diego Zapata-Rivera, Carolyn M. Forsyth, & Christine Luce

Educational Testing Service, Princeton, NJ

Substantial progress has been made toward applying technology enhanced conversation-based assessments (CBAs) to measure the English-language proficiency of English learners (ELs). CBAs are conversation-based systems that use conversations among computer-animated agents and a test taker. We expanded the design and capability of prior conversation-based instructional and assessment systems and developed a CBA designed to measure the English language skills and the mathematics knowledge of middle school ELs. The prototype CBA simulates an authentic and engaging mathematics classroom where the test taker interacts with two virtual agents to solve math problems. We embedded feedback and supports that are triggered by how the CBA interprets students' written responses. In this study, we administered the CBA to middle school ELs ($N = 82$) residing in the United States. We examined the extent to which the CBA system was able to consistently interpret the students' responses (722 responses for the 82 students). The study findings helped us to understand the factors that affect the accuracy of the CBA system's interpretations and shed light on how to improve CBA systems that incorporate scaffolding.

Keywords Conversation-based assessments; conversation path; accuracy; scaffolding; English learners

doi:10.1002/ets2.12315

Over the last few years, substantial progress has been made in exploring the capabilities of applying technology-enhanced conversation-based tasks facilitated by virtual agents to measure the English-language proficiency (ELP) of English learners (ELs; Lopez & Guzman-Orth, 2015; So et al., 2015). The capability of a conversation-based assessment (CBA) consists of the ability to design, implement, and score naturalistic, communicative, interactive tasks that simulate learning strategies, processes, and social interactions. Virtual agents have become prevalent in the last decade in online games and in learning environments, and virtual agents within a CBA can be programmed to have different personalities and functions, such as virtual tutors, classmates, or peers in the subject matter being assessed (Graesser et al., 2017). We have been exploring the use of virtual agents in ELP assessments by combining the use of agents with speech recognition and language processing systems and by using tasks involving multiple agents and a human test taker to create more realistic assessment environments. By using CBAs, it is possible to create naturalistic communicative tasks that can include dynamic conversations and scaffolded feedback. CBAs have the potential to allow for elicitation of more detailed information about EL students' language skills and mathematics knowledge because CBAs are interactive and engaging, provide immediate feedback, and include additional prompting and scaffolds (Lopez et al., 2017).

Conversation-Based Systems

Graesser et al. (2014) described conversation-based systems as conversations among computer-animated agents and a human student. In conversation-based systems, tasks are structured to provide opportunities not only for individuals to give evidence of their knowledge, skills, and abilities, but also for instructional designers to scaffold learning and provide useful feedback to students. One of the unique characteristics of conversation-based systems is that through the natural flow of conversation, virtual agents can ask follow-up questions and provide hints to elicit additional or missing information. Virtual agents can also repeat or rephrase their questions and provide feedback on the quality of the responses. Technology-enhanced conversation systems involving virtual agents and a human student have been successfully used

Corresponding author: A. A. Lopez, E-mail: alopez@ets.org

as part of intelligent tutoring systems to facilitate learning (Chan & Baskin, 1990; Graesser et al., 2001; Hong et al., 2014; Johnson et al., 2000; Yang & Zapata-Rivera, 2010).

Technology-enhanced conversation systems have also been used for assessment purposes. CBAs are automated conversations that involve dialogs between virtual agents and test takers to measure knowledge and skills and have been utilized in multiple domains (Zapata-Rivera et al., 2015). For example, CBAs have been used to measure argumentation skills (Song & Sparks, 2017), science inquiry (Zapata-Rivera et al., 2014), mathematical argumentation (Cayton-Hodges, 2016), and ELP skills (Evanini et al., 2014; Mitchell et al., 2014; So et al., 2015). In the next sections we focus on two themes, conversation-based systems in language learning and conversation-based systems in language assessment.

Background

Conversation-Based Systems in Language Learning

Conversation-based tasks create learning environments that simulate particular pedagogical strategies (Butler et al., 2011). These strategies include cooperative learning, inquiry-based learning, visualization, interactive learning, and scaffolding. All these strategies are particularly effective for language learning. Consequently, conversation-based tasks have been used to help facilitate second language learning. Yang and Zapata-Rivera (2010) developed a conversation-based pragmatics game called *The Request Game* to facilitate second language pragmatic learning by providing learners an alternative to practice communicative skills in a stress-free, engaging environment. Using a dialog engine and a virtual agent, language learners practiced pragmatics by making requests to the virtual agent in simulated interactions. The virtual agent provided both verbal and non-verbal (facial expressions) feedback based on the appropriateness of the request entered by the human learner. *The Request Game* was well received by users as an engaging and easy way to practice pragmatics. Yang and Zapata-Rivera also reported that the dialog engine's ability to process complex language improved by gathering more examples of how students responded to the questions.

Hong et al. (2014) implemented computer-animated agents as an instruction tool to facilitate social interaction and student engagement for elementary-aged Taiwanese students by developing courseware, an xml-based authoring tool, for use as part of the classroom curriculum to support English learning. Incorporating on-screen text, images, video, and audio files, the virtual agent performed as a tutor for the students and as a teaching assistant for the instructors. Hong et al. conducted an experiment using two groups of elementary school-aged English learners in Taiwan. The experimental group used the virtual agent scripted materials while the control group used traditional classroom materials. Results revealed students who interacted with the virtual agent were "more actively engaged" in the learning activities and outperformed students in the control group. Students stated they felt at ease interacting with the virtual agent, who made learning English "joyful." Students also stated they would like to speak with two or more virtual agents and hoped to use virtual agents for learning in their other classes. Instructors commented positively on the ability of the virtual agents to pronounce English words and suggested the courseware could serve as a way for the virtual agents to assist in teaching the classroom lessons in a realistic way, increasing students' attention and engagement.

Conversation-Based Systems in Language Assessment

The potential to expand instructional integrated conversation-based systems for assessment purposes was investigated by So et al. (2015). The purpose of their investigation was to determine if a CBA could be used to measure the ELP of young ELs in Grades 3 – 5. The CBA developed by So et al. required a human student to interact with two virtual agents in an interactive scenario-based environment (classroom and school library). The findings from this study provided evidence to support using CBAs to measure the English language skills of second language learners in a realistic, engaging environment. Although most students reported positive feedback about interacting with the virtual agents, several students did not like the appearance of the virtual agents, who had a cartoon-like appearance and thus did not realistically resemble middle school students. Data collected by So et al. also provided valuable speech samples to train and refine the system for future studies.

Evaluation of Response Quality in CBAs

Evanini et al. (2014) investigated the possibility of incorporating an automated speech recognition (ASR) system into So et al.'s (2015) CBA to extend the system to process spoken responses from the human test taker. Students were required to listen and respond to the agents through a series of interactive classroom tasks and enter both a written response and a spoken response. The spoken responses were recorded to train the ASR for future use, while the written responses determined the virtual agent's next dialog move. Evanini et al. continued development on the ASR, and later versions of the CBA included the ASR system implemented into the entire CBA prototype, eliminating the need for students to enter a written text response to advance to the virtual agent's next dialog. Forsyth et al. (2019) also conducted research using the same CBA prototype to compare how students responded to virtual characters and humans. Their research suggested that students' responses to agents are like those provided by humans in summary explanations in CBA, but they have not yet investigated these responses in relation to feedback.

Task Design

An interactive prototype CBA system, the ELLA-Math CBA, was designed to measure the ELP and math content knowledge of EL middle school students. The ELLA-Math CBA is intended to be an authentic representation of small-group work activities in a middle-school mathematics classroom where the test taker interacts with three virtual agents: a teacher (Ms. Davidson) and two student agents (Sara and Lucas). However, most of the conversations in this task occur between the student and the two student agents. Figure 1 shows a snapshot of the task interface with the three virtual agents. The task was designed from the perspective of the test taker as an active participant in the small group, meaning that no physical entity represents the test taker in the environment, similar to the first-person perspective used in many video games.

The ELLA-Math CBA includes six conversations that guide the student through a math worksheet given to them by their teacher. Three of these conversations start with a multiple-choice question; the others start with an open-ended question. The multiple-choice question facilitates the design of the conversation task by limiting the context of the conversation. Moreover, multiple-choice questions allow the inclusion of previously identified topics or misconceptions in the options or distractors based on the review of the literature or based on students' responses in prior studies. In this study, we focused only on the three conversations that start with an open-ended question because we wanted to examine if the CBA system could consistently interpret the students' written responses and to understand the factors that affect the system's accuracy in interpreting their written responses. Each of these conversations starts with a main question and the interaction flows naturally within and across conversations to simulate authentic turn-taking behaviors between the virtual agents and the student. The students participate in the conversations by typing their response in the computer interface. In the Following Directions conversation, the teacher gives directions about the assignment, while in the other two conversations, the students work in groups. In the Ratios conversation, the students and the virtual agents answer a question about ratios, and in the Unit Rate conversation, they answer a question about unit rate.

The CBA system attempts to understand the students' written responses during the interaction with the virtual agents using computer natural language processing (e.g., regular expressions and latent semantic analysis [LSA]) to match the students' answer to an expected correct response to the given question producing a match score ranging from 0 to 1. Specifically, regular expressions (Jurafsky & Martin, 2008) focus on creating a match score based on a single word or phrase. LSA (Landauer et al., 2007) analyzed student input based on a geometric pattern-matching algorithm attempting to understand the meaning of the input based on its reference corpus (i.e., real world input). In the given context, both algorithms are considered and a threshold for this match score is determined based on iterative testing to decide when an answer can be deemed correct. This approach for natural language processing has been successful and comparable to human raters in matching student input to expected responses in intelligent tutoring systems (Cai et al., 2011).

Based on the interpretation of the student input, the system will designate the next level of scaffolding or none at all, depending on the quality of each student's response (i.e., match score). Thus, students can be sent down a variety of conversational paths based on their match score(s) (see Figure 2). For example, a student's response can be correct (correct conversation path), partially correct (partial conversation path), or incorrect (incorrect conversation path). Sometimes the system is unable to decide because it cannot interpret what a student types (insufficient conversation path). This situation happens when the student says something irrelevant (off topic, not relevant to the topic of the question), requests

Figure 1 Screenshot of virtual agents in the CBA math classroom. CBA = conversation-based assessment.

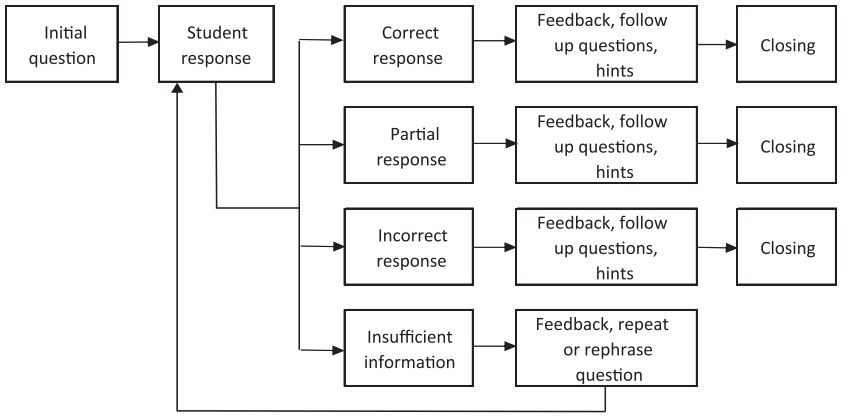

Figure 2 Sample conversation diagram.

clarification (e.g., I do not know, please repeat), provides unexpected answers that might be on topic but are not categorized by the current natural language processing, or does not provide a response. In this case, the system will prompt the student by giving them a second opportunity to respond.

In each conversation path, the virtual agents provide feedback on the quality of the response, and support (e.g., hints, follow-up questions) in order to allow the student to explain his or her responses or give more information. The feedback and supports are embedded in the conversations and are triggered by how the CBA interprets the student's response.

Purpose of the Study

We embedded scaffolding (i.e., feedback and supports) that is triggered by how the CBA interprets the students' written responses. With this addition, the CBA considers the entire conversation path, not only the responses to the opening question, to measure the ELP and math content knowledge of EL students. Therefore, it is important to manage the conversation by interpreting whatever the student says or types and responding appropriately to each response. To manage the conversation well, the virtual agents' actions, the feedback and hints they provide, and the follow-up questions they ask must be relevant and appropriate. If student responses are not interpreted accurately and students are sent through an inappropriate conversation path, the resulting information about students' skills and knowledge may be invalid (i.e., scores will be assigned on the basis of an inappropriate conversation path). Thus, it is critical to examine how accurate these types of CBA systems interpret the students' responses.

The purpose of the study is to examine how accurately the ELLA-Math CBA system interprets EL students' responses and determines which conversation path the students need to follow (which conversational paths to take) and to find ways to improve the CBA's ability to accurately interpret student responses. Three major research questions guided our study:

1. How accurate is the prototype ELLA-Math CBA system in interpreting student responses?
2. What difficulties does the prototype ELLA-Math CBA system have in interpreting student responses?
3. How do students perceive the accuracy of the prototype ELLA-Math CBA system in interpreting their responses?

Method

In this section, we provide information about the participants, data sources, data collection procedures, and data analyses.

Participants

Participants for the study were recruited from two urban school districts in the southern and eastern United States. Inclusion criteria required that students were ELs enrolled in Grades 6, 7, or 8. The final participant count was 82 EL students. Across school sites, the breakdown was as follows: Site 1 had 73 EL students and Site 2 had nine. Out of the 82 total students, 39 students were male and 43 were females. There were 46 students in Grade 6, 23 in Grade 7, and 13 in Grade 8. Their ages ranged from 11 to 15. The majority of the students were U.S.-born EL students (57). The rest of the students were born in other countries: Mexico (7), El Salvador (3), Honduras (3), India (3), Ecuador (2), Colombia (1), Dominican Republic (1), and Spain (3). Three students did not report their country of origin.

Data Sources and Procedure

We investigated a data set with 242 responses to three initial questions and 480 responses to nine follow-up questions. In total, we included 722 responses for the 82 students. We also designed a survey with 20 questions to gauge students' perceptions regarding the prototype CBA system (e.g., the authenticity of the environment and their experience interacting with the virtual agents). For the purpose of this study, we focused only on the four questions related to the students' perceptions of the accuracy of the CBA system in interpreting their responses (i.e., if they understood what the virtual agents said, if the virtual agents said things that made sense, if the follow-up questions were useful, and if the feedback and supports were helpful). We also asked teachers to complete a questionnaire to get background information about the EL students (e.g., gender, age, grade level, country of origin, length of time in the United States, and languages spoken at home). The CBA was administered on the computer using Internet Explorer in a 90-min class session. Participants began by completing the CBA. Upon completion of the conversations, students then were asked to complete the survey.

Data Analyses

The data analyzed for this study included the students' responses and scores to each question in the ELLA-Math CBA, information about students from the background questionnaire, and responses on the survey. We used path-based scoring in this study, meaning that all the responses in the ELLA-Math CBA were scored automatically according to the virtual agent's dialog move (the conversation path that was assigned). We then extracted all data from the participant log files. The data were cleaned and reviewed, with care taken to ensure any student misspellings or other keyboarding errors remained intact and were not changed. All errors (misspellings, grammar, syntax, or other keyboarding errors) are those of the student and were retained for analysis. To evaluate the accuracy of the CBA system in interpreting students' responses, we compared the CBA system's path-based scores to judgments of experts (Graesser et al., 2000). Two of the researchers scored each response independently, and then held a discussion to compare their scores and reach a consensus score for each response. We then examined if the scores assigned by the CBA system matched the consensus scores assigned by the raters. We added the number of times there was a match and then divided by the total number of responses to get an accuracy rate. We also analyzed qualitatively all the responses that were not interpreted accurately by the system to identify patterns. Next, we analyzed the rest of the responses to determine if the same patterns were also present. Additionally, to summarize the students' perceptions about their experience interacting with the virtual agents, we analyzed their responses to obtain descriptive statistics (frequency and percentage).

Table 1 Accuracy Rate for Interpreting Responses to Initial Questions

Question	Path	Expected responses	Number of responses	Interpreted correctly by the CBA	Accuracy rate (%)
What did Ms. Davidson asked us to do in class? Following Directions conversation ($N = 82$)	Correct	Solve math problems and work in groups	11	11	100
	Partial	Solve math problems or work in groups	38	26	68
	Incorrect	Other things the teacher said (e.g., say name)	8	6	75
	Insufficient	No responses, clarification requests, off topic or ambiguous responses	25	24	96
What is the ratio of cups of raisins to grams of fiber? Ratios conversation ($N = 80$)	Correct	1:10, 1/10, 1 – 10, or one to ten	13	13	100
	Incorrect	Any other ratio or a number (e.g., 10:1, 10)	41	39	95
	Insufficient	Same as above	26	25	96
What is the unit rate for the number of grams of fiber per cup of broccoli? Unit Rate conversation ($N = 80$)	Correct	3:1	4	4	100
	Partial	4:12	4	4	100
	Incorrect	Any other number or ratio	66	61	92
	Insufficient	Same as above	6	6	100

Note. CBA = conversation-based assessment.

Results

The results section is split into four sections: accuracy in initial constructed-response questions, accuracy in follow-up questions, sample student responses that were interpreted accurately or misinterpreted by the CBA system for all three conversations, and students' perceptions on the accuracy of the CBA system in interpreting their responses. In this study, accuracy refers to how well the scores assigned by the CBA system matched the consensus scores assigned by the raters.

Accuracy in Interpreting Responses in Initial Open-Ended Questions

Table 1 provides information about the initial open-ended question in each of the three conversations, the expected responses for each conversation path, and the accuracy rate of the CBA system in interpreting responses to the initial questions. The CBA system was highly accurate in interpreting responses to the three initial open-ended questions; 219 of the 242 responses were interpreted accurately. The CBA system was more accurate interpreting responses in the initial math-related questions (96.3% in the Ratios conversation and 93.8% in the Unit Rate conversation) than in the initial language-related question (81.7% in the Following Directions conversation). In the Ratios conversation, three of the 80 responses were not interpreted accurately whereas in the Unit Rate conversation, five of the 80 responses were not interpreted accurately. In the Following Directions conversation, the CBA system misinterpreted 15 of the 82 responses: 12 were sent though the partial conversation path, two through the incorrect conversation path, and one through the insufficient conversation path. The system accurately interpreted 100% of the responses that were sent through the correct conversation path: 11 responses in the Following Directions conversation, 13 in the Ratios conversation, and four in the Unit Rate conversation.

Accuracy in Interpreting Responses in Follow-up Questions

Table 2 provides information about the follow-up questions in each of the three conversations, the expected responses for each conversation path, and the accuracy rate of the CBA system in interpreting the follow-up responses. The system was very accurate in interpreting the students' responses in all the follow-up questions: 439 of 480 follow-up responses were interpreted accurately by the system (91.5%). Similar to our findings on the responses to the initial questions, the

system accurately interpreted 100% of the follow-up responses that were sent through the correct conversation path. In the Following Directions conversation, the total accuracy rate was 97.4%. All the responses to the first follow-up question were interpreted accurately whereas one response was misinterpreted in the second follow-up question. This correct response was sent through the incorrect conversation path instead.

The total accuracy in interpreting the responses to the follow-up questions in the Ratios conversation was 90.4% (83.6% in the first follow-up question and 92.8% in each of the other three follow-up questions). In the first follow-up question, 10 of the 61 responses were not interpreted accurately and were all sent through the insufficient conversation path. In each of the other three follow-up questions, four responses were not interpreted accurately. All these responses were sent through the incorrect conversation path. In the Unit Rate conversation, all the responses to the second and third follow-up questions were interpreted accurately. Five responses to the first follow-up question were not interpreted accurately. These responses were sent through the incorrect conversation path.

Interpreting Student Responses in the Following Directions Conversation

The CBA system was able to interpret accurately some functional synonyms (i.e., expressions that convey similar meaning in the context of a conversation) in the responses to the initial question in the Following Directions conversation (see Table 2). For example, the CBA system was able to interpret that other expressions conveyed the meaning that the students had to *solve math problems*. Thus, responses such as "That we would answer some math questions," "Figure out some math problems involving food," or "you both need to take turns reading the problems and i will write them down" were interpreted accurately and sent through the appropriate conversational path (underlining added for emphasis). Likewise, the CBA system interpreted accurately other expressions that conveyed the meaning that the students had to *work in groups*. Thus, responses such as "help each other with math problems about food," "you both need to take turns reading the problems and i will write them down," or "we need to work as a group and answer some math problems" were interpreted accurately. The system also interpreted correctly when students described their roles, (e.g., "you read the problem, I write down answers"), to convey that they had to work in groups. However, the system was not able to recognize other functional synonyms that conveyed the idea that the students had to work in groups. Even though the system included listing the student's role (e.g., "you read, I type") to convey that they had to work together, the system did not include *to record* as one of the roles the students had. Consequently, a response like "to go over the question and i will record the answer" was sent through the partial conversational path instead of the correct conversational path.

The CBA system was also able to interpret appropriately a few misspellings. For example, the CBA system accurately interpreted one of the responses even though the word *problems* was misspelled: "she said we are going to be doing math problomes you two are going to take turns reading the question." However, in most instances, the CBA system misinterpreted correct responses because students misspelled a few key words in the expected response. For example, key words such as *take turns*, *read*, *answer*, *problems*, *assignment*, and *questions* were misspelled in the Following Directions conversation. Below are a few examples (underlining added for emphasis):

- "each other are going to take terns reading the work"
- "you and the boy have to red the math questions and i will write the anserw"
- "to solve math problens and you both will take turns to read"
- "we have to solve promblems with each other"
- "to do math assigment all of us together and discuse it."
- "That we were going to ask some math probles about food."
- "That we had to do some math qutions and some words to know"
- "we were going to do math prorloms with food"

Interpreting Student Responses in the Ratios Conversation

The CBA system was able to interpret accurately all the ratios in the responses to the initial question in the Ratios conversation regardless of the notation that was used (see Table 2). For example, the CBA system interpreted different notations for the ratio *one to ten*. Some students wrote "1:10"; others wrote either "1/10," or "1 to 10." Even more elaborate responses

Table 2 Accuracy Rate for Interpreting Responses to Follow-Up Questions

Conversation	Follow-up question	Path	Expected responses	Number of responses	Interpreted correctly by the CBA	Accuracy rate (%)
Following Directions	Do we have to work in groups or by ourselves? ($N = 30$)	Correct	In groups	30	30	100
		Incorrect	Any other response (e.g., alone)	0	—	—
		Insufficient	No responses, clarification requests, off topic or ambiguous responses	0	—	—
	What exactly do we have to do in groups ($N = 8$)	Correct	Solve math problems	5	5	100
		Incorrect	Any other response (e.g., discuss words)	1	0	0
		Insufficient	Same as above	2	2	100
Ratios	What is the problem asking us to do? ($N = 61$)	Correct	Ratio of cups of raisins and grams of fiber	5	5	100
		Partial	Find ratio	2	2	100
		Incorrect	Any other response	1	1	100
		Insufficient	Same as above	53	43	81
	Do you agree or disagree with Lucas? ($N = 56$)	Correct	Agree	12	12	100
		Incorrect	Disagree	31	27	87
		Insufficient	Same as above	13	13	100
	How many grams of fiber are in one cup of raisins? ($N = 56$)	Correct	10	12	12	100
		Incorrect	Any other number	31	27	87
		Insufficient	Same as above	13	13	100
	Write this as a ratio? ($N = 56$)	Correct	1:10, 1/10 or 1 to 10	12	12	100
		Incorrect	Any other ratio	31	27	87
		Insufficient	Same as above	13	13	100
Unit Rate	What is the problem asking us to do? ($N = 71$)	Correct	Find unit rate	26	26	100
		Incorrect	Any other response (e.g., find cups of broccoli)	15	5	33
		Insufficient	Same as above	30	22	73
	How many grams of fiber are in four cups of broccoli? ($N = 71$)	Correct	12	50	50	100
		Incorrect	Any other number	8	8	100
		Insufficient	Same as above	13	13	100
	How many grams of fiber are in one cup of broccoli? ($N = 71$)	Correct	3	51	51	100
		Incorrect	Any other number	19	19	100
		Insufficient	Same as above	1	1	100

Note. CBA = conversation-based assessment.

were also interpreted as correct (e.g., "1 cup of raisins to 10 grams of fiber" or "the ratio of the number of grams of fiber is 1:10"). Ratios other than "1:10" were interpreted properly as incorrect responses (e.g., "10," "10 to 20," "1:03," or "1 of 10").

However, three of the responses to the initial question in this conversation were not interpreted accurately. In one case, a student wrote "1to10." This response should have been interpreted as correct, but the system was not able to recognize the lack of spacing between the numbers and the word *to*. In another case, another student wrote "1 over 10." In this case, we think this student wanted to write the ratio "one to ten" using the notation $\frac{1}{10}$, but was not able to find the appropriate symbol on the keyboard, so this response could have been interpreted as correct. Similarly, one of the students wrote "1 oz./10 g," but this response was incorrectly interpreted as an irrelevant response because the abbreviations for ounces and grams were not taken into consideration in the expected responses for the correct conversational path.

In the first follow-up question in the Ratios conversation, 12 responses were not interpreted accurately. The main issue was that students misspelled key words (e.g., *ratio*, *raisins*, *fiber*). So responses such as "to do the rayio of number of cups of raisins over the number of grams of fiber," "what the ratio of cup of raisinc and grams of fiber" and "to figure out the ratio on the cups of raisins to the number of grams of fiver" (underlining added for emphasis) were not interpreted accurately by the system. Similarly, four responses were misinterpreted in the second follow-up question in this conversation. In this question, students were asked whether they agreed or disagreed with Lucas's definition of ratio. One of the students misspelled the word *agree*, and the other three students misspelled the word *disagree* ("yes i agre," "disagree," "disigre," and "I think i desagree"). The CBA system misinterpreted these responses as insufficient and asked the students to respond again. The first response should have been sent through the correct conversational path; the other three should have been sent through the incorrect conversational path.

The system also had problems interpreting four correct responses in the third follow-up question. The expected correct response was *10*, but the students wrote two numbers in their responses ("in one cup there is 10 grams of fiber," "There are 10 grams of fiber in 1 cup of raisins," "1 cup of raisins =10," and "10 grams = 1 cup of raisins"). The system misinterpreted these responses, and they were sent through an inappropriate conversational path. In the last follow-up question, students were asked to write a ratio. The CBA system had similar difficulties in interpreting the lack of spacing (e.g., "1Cup is to 10 Fibers") and different ratio notations (e.g., "1 over 10" and "We put the 1 on top of the 10").

Interpreting Student Responses in the Unit Rate Conversation

In the initial question for the Unit Rate conversation, students were asked to write the unit rate of grams of fiber per cup of broccoli (see Table 2). The CBA system accurately interpreted different notations for the ratio *three to one* in this conversation (e.g., "3 to 1" or "3:1") and sent the responses through the correct response path. Equivalent ratios (ratios that express the same relationship between numbers) were sent accurately through the partial response path (e.g., "12/4," "12:4," "12 is to 4 cups"). Other ratios and numbers were interpreted appropriately and sent to the incorrect response path (e.g., "12 grams," "1:03," "1:4," or "48"). Only five of the 66 responses that were sent through the incorrect conversational path were not interpreted accurately. Instead of writing a ratio, these students wrote responses such as "3 grams of fiber per each cup," and "3 grams of fiber per cup of broccoli." The problem was the CBA system was not able to determine that the number of cups was implied in these responses. In both of these examples, it is implied that the number of grams of fiber in 1 cup of broccoli is three.

All the responses to the follow-up questions in this conversation were interpreted accurately except for 10 of the 15 responses that were sent through the incorrect conversational path in the first follow-up question. This question asked students to explain what the problem was asking them to do. The problem is that the expected correct response for this question was too restricted and required students to include the phrase *unit rate* in the response. So responses such as "to find the number of grams of fiber per cup of broccoli," and "to find how much gram of fiber are in one cup of broccoli" were not interpreted accurately as correct responses.

Students' Perceptions about the Accuracy of the CBA System

We also asked students to report their perceptions about interacting with the virtual agents (see Table 3). Almost half of the students reported that the two virtual agents did not always understand their responses (49.4%). For instance, a student commented that she "didn't like when they said they didn't understand me" (Grade 6, female student). Students would often get frustrated whenever they felt the CBA system did not understand their responses. One student stated that

Table 3 Student Perceptions on Interacting With the Virtual Agents

Statement	*N*	Strongly disagree	Disagree	Agree	Strongly agree
I felt that Sara and Lucas always understood my responses	79	12 (15.2%)	27 (34.2%)	31 (39.2%)	9 (11.4%)
Sara and Lucas said things that did not make sense to me	77	14 (18.2%)	31 (40.3%)	26 (33.8%)	6 (7.8%)
The extra questions from Sara and Lucas helped me answer the main problems	78	5 (6.4%)	10 (12.8%)	44 (56.4%)	19 (24.4%)
The comments and hints that Sara and Lucas gave helped me answer the main problems	79	8 (10.1%)	7 (8.9%)	50 (63.3%)	14 (17.7%)

the part of the activity he did not like was "When they [Sarah and Lucas] say they don't understand me because it annoys me" (Grade 7, male student). Also, 41.6% of the students reported that the two virtual agents sometimes said things that did not make sense to the students. Specifically, the students reported that the virtual agents were saying things the student had already stated. For example, some students commented that sometimes the virtual agents were not understanding the student responses because the agents were asking the students to elaborate or be more specific even though the students felt they had already done so. One of the students explained that the two virtual agents "asked me questions and I responded but they said that I had to be more specific so I gave a more specific response, but their response was basically the same as my first response."

Discussion

Our first research question sought to determine how accurate the CBA system was in interpreting the students' responses. We found that, for the most part, the system was able to consistently match the students' responses to the preassigned expectations in each of the questions and send them through an appropriate conversation path. The system was able to recognize some words or phrases that conveyed the same meaning as the expected correct responses (i.e., functional synonyms). Some responses were interpreted accurately even when a few key words were misspelled. It is important to highlight that the system was always accurate in interpreting the responses that were sent through the correct response conversation path. That is, we did not find any instances in which an incorrect or partial response was misinterpreted as a correct response. However, there were a few instances where some students wrote correct responses, but the system sent them through other conversation paths (e.g., partial, incorrect, or insufficient conversation path). Because each conversation path includes follow-up questions, hints, and feedback, some of these students were able to provide additional information or the correct response.

We also found that the CBA system was more accurate in interpreting responses to questions that expected students to write a number than in interpreting responses that required students to write words (e.g., a single word, phrase, or sentence). Likewise, the system interpreted shorter responses (e.g., numbers, one- or two-word responses) more accurately than longer responses (e.g., phrase or sentence). The system was also more accurate when the expected responses were more flexible and accepted multiple functional synonyms (e.g., work in groups, work together, take turns) than when the expected response required students to include a specific word or words in the response (e.g., ratio, cups of raisins, grams of fiber).

The second question examined the problems the CBA system had in interpreting student responses. The most common problem was that the students' responses had a high incidence of misspellings and typographical errors. Some of the typing errors included misspelling of key words in the expected correct responses for each of the conversations. Although the system was able to interpret a few of them as correct, sometimes the system sent responses that included misspelled words through the partial, incorrect, or insufficient conversation paths. When this occurs, scores might be affected, and students might get follow-up questions, hints or feedback that are not appropriate or relevant. Moreover, a few students had difficulties with keyboarding issues. Either they were not able to find a specific key, or they had problems with spacing. Although a few students had difficulties typing math symbols, they used their entire linguistic repertoire to find other

ways to express their ideas. Some studies have found that students tend to use words instead of symbols when they cannot find the desired symbols on a traditional QWERTY keyboard (e.g., Lopez et al., 2019). Many ELs are still developing their writing skills and some may lack enough keyboarding skills; thus, their written responses will typically include misspellings, typos, and grammatical and typographical errors (e.g., capitalization, punctuation, spacing). CBA systems need to be sensitive to EL students' typing errors.

Another problem we identified was that the system did not recognize some words or expressions that conveyed the same meaning as the expected correct response (i.e., functional synonyms). Although some questions accepted multiple functional synonyms (e.g., work in groups, work together, take turns) other questions required students to include a specific word in a response (e.g., ratio). The CBA system was more accurate when multiple functional synonyms were included in the expected responses. Thus, efforts should be made to include as many functional synonyms as possible. Some can be anticipated during the test development phase, and others can be added after testing the system with actual students. For example, in the first follow-up question in Conversation 2, students were required to use the word "ratio" to explain what the problem was asking them to do (e.g., "find the ratio between the cups of raisins and the grams of fiber."). Other responses, such as "find the relationship between the cups of raisins and the amount of grams of fiber," could also be included in the expected response.

The third question examined how students perceived their interaction with the virtual agents and the accuracy of the CBA system in interpreting their responses. We found that, for the most part, students felt the system was able to respond appropriately to what they were saying and that the agents' responses were relevant. They also liked that the agents helped them understand the questions by asking follow-up questions and providing hints and feedback. However, some students reported that the agents did not always understand their responses. In some cases, students felt they had already provided the information the agents were requesting. Many of these students became frustrated every time they felt the agents were not understanding them. This sense of frustration was also noted in the student responses when students used capital letters and/or added exclamation marks to answer the follow-up questions (e.g., "*TO FIND THE GRAMS OF FIBER OF THE CUPS OF RAISIN!!!!!*).

Recommendations for Improving the Accuracy of CBA Systems

Because we cannot anticipate all the different responses students may provide, we cannot design CBAs to be completely accurate under the current approach. However, there are ways to improve them. During the design and development phase, test designers can ask experts to revise and expand the list of expectations for each initial and follow-up question. The accuracy of the CBAs can also be evaluated during the design and development phase. Although evaluation can be done manually, this process requires significant time and effort. We recommend the use of automated testing tools to speed-up the process (Zapata-Rivera et al., 2015). Automated testing of CBAs requires creating or gathering sample responses from experts (e.g., correct, partial, incorrect, or irrelevant responses) and evaluating them to determine if the CBA system can process them accurately. This information can be used to refine the expectations. The CBAs also need to be piloted with students. The students' responses can be used to examine the accuracy of the CBAs and to refine the expectations in each of the conversational paths in the CBA. Data collected can be used to improve the regular expressions used to evaluate expressions in the systems. Different approaches to make the system more robust are being explored, including hybrid approaches that use both machine learning and input from human experts (Zapata-Rivera et al., 2019). A major challenge is that many ELs are still developing their English skills, so it is important to develop CBAs that can effectively accommodate many forms of misspelled words and typographical errors. Students' misspellings and typographical errors can be handled by integrating a spelling module that corrects misspellings before the response is evaluated by the system's use of regular expressions and latent sematic analysis.

Limitations and Future Research Directions

This study was conducted in only two schools in two states. Therefore, the findings may have somewhat limited generalizability. Also, the sample size was relatively small; only 82 students participated in this study. That said, we did analyze 722 responses from multiple questions, three initial questions and nine follow-up questions for the 82 students who participated in the study. Focusing on a small sample is also consistent with standard methods for instrument development and

piloting. Another limitation in this study is that most of the participants shared the same language background (i.e., Spanish). Thus, some of the problems the system had in interpreting the student responses might be related to their language background. Despite these limitations, the study findings provide valuable information on understanding what factors affect the accuracy of a CBA designed for EL students and on how to improve its accuracy.

Although the use of CBA is promising, more empirical evidence supporting its validity is needed. For example, it is critical to examine what level of accuracy rate is adequate for the purpose of a CBA and how its accuracy rate impacts the type of feedback students get. We also found that some students expressed frustration with the type of feedback they were receiving so it is important to examine the level of accuracy rate needed to minimize students' levels of frustration. Equally important is to address some fairness issues. For instance, future studies could examine the impact that EL students' spelling skills and keyboarding ability have on scores and feedback.

Further research can build on the present study with a more diverse sample of students in terms of terms of their ELP, educational background, math knowledge, grade level, home language, experience taking mathematics assessments, and length of time in the United States. Also, we only examined the accuracy of a CBA system in interpreting written responses. Therefore, study findings can be generalizable only to written, not oral responses. Future studies could also examine the accuracy of CBA systems that include oral responses.

Conclusions

Overall, we found that the ELLA-Math CBA system was highly accurate in interpreting EL students' written responses, and for the most part, it was able to send students through the appropriate conversation paths. In most cases, the feedback, hints, and follow-up questions allowed EL students to explain their initial responses or to give more information to accurately assess their language proficiency and mathematics knowledge. The findings from this study also provided useful information about the difficulties the ELLA-Math CBA had in interpreting the students' written responses. In turn, we can use this information to improve CBA systems that incorporate scaffolding and multiple conversation paths that are triggered by the students' responses so they can be interpreted and scored accurately, and that the virtual agents can respond in an appropriate and relevant manner. Finally, by improving the accuracy of this type of CBA systems, we can also lessen the frustration level of the students while they interact with the virtual agents.

References

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, E. Amsel, B. M. Kowalski, B. C. Beins, K. D. Keith, & B. F. Peden (Eds.), *Promoting student engagement: Volume 1. Programs, techniques and opportunities* (pp. 286–291). Society for the Teaching of Psychology. http://teachpsych.org/ebooks/pse2011/vol1/index.php

Cai, Z., Graesser, A. C., Forsyth, C., Burkett, C., Millis, K., Wallace, P., Halpern, D., & Butler, H. (2011). Trialog in ARIES: User input assessment in an intelligent tutoring system. In W. Chen & S. Li (Eds.), *Proceedings of the 3rd IEEE International Conference on Intelligent Computing and Intelligent Systems* (pp. 429–433). IEEE Press.

Cayton-Hodges, G. A. (2016). Assessing mathematical argumentation through automated conversation. In M. B. Wood, E. E. Turner, M. Civil, & J. A. Eli (Eds.), *Proceedings of the 38th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 1471–1478). University of Arizona.

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). Ablex Publishing.

Evanini, K., So, Y., Tao, J., Zapata-Rivera, D., Luce, C., Battistini, L., & Wang, X. (2014). Performance of a trialogue-based prototype system for English language assessment for young learners. In K. Berkling (Chair), *Fourth workshop on child computer interaction* (pp. 79–84). ISCA. https://www.isca-speech.org/archive/wocci_2014/papers/wc14_079.pdf

Forsyth, C. M., Luce, C., Zapata-Rivera, D., Jackson, J., Evanini, K., & So, Y. (2019). Evaluating English language learners' conversations: Man vs. machine. *International Journal on Computer Assisted Language Learning*, *32*(4), 398–417. https://doi.org/10.1080/09588221.2018.1517126

Graesser, A. C., Forsyth, C. M., & Lehman, B. A. (2017). Two heads may be better than one: Learning from computer agents in conversational trialogues. *Teachers College Record*, *119*(3), 1–20.

Graesser, A. C., Li, H., & Forsyth, C. (2014). Learning by communicating in natural language with conversational agents. Current directions. *Psychological Science*, *23*(5), 374–380. https://doi.org/10.1177/0963721414540680

Graesser, A. C., Person, N., & Harter, D. (2001). Teaching tactics and dialog in AutoTutor. *International Journal of Artificial Intelligence in Education*, *12*(3), 257–279.

Graesser, A. C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., & Person, N. (2000). Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments*, *8*(2), 129–147. https://doi.org/10.1076/1049-4820(200008)8:2;1-B;FT129

Hong, Z. W., Chen, Y. L., & Lan, C. H. (2014). A courseware to script animated pedagogical agents in instructional material for elementary students in English education. *Computer Assisted Language Learning*, *27*(5), 379–394. https://doi.org/10.1080/09588221.2012.733712

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*(1), 47–78.

Jurafsky, D., & Martin, J. (2008). *Speech and language processing*. Prentice Hall.

Landauer, T., McNamara, D. S., Dennis, S., & Kintsch, W. (2007). *Handbook of latent semantic analysis*. Erlbaum. https://doi.org/10.4324/9780203936399

Lopez, A. A., & Guzman-Orth, D. (2015, April 16–20). *Using conversation–based assessments to diagnose the English language skills and math knowledge of middle school English learners* (Paper presentation). AERA. Chicago, IL.

Lopez, A. A., Guzman-Orth, D., & Turkan, S. (2019). Exploring the use of translanguaging to measure the mathematics knowledge of emergent bilingual students. *Translation and Translanguaging in Multilingual Contexts*, *5*(2), 143–164. https://doi.org/10.1075/ttmc.00029.lop

Lopez, A. A., Luce, C., Zapata-Rivera, D., & Forsyth, C. (2017). Using formative conversation-based assessments to support students' English language development. *IEEE Technical Committee on Learning Technology Bulletin*, *19*(1), 6–9. https://tc.computer.org/tclt/wp-content/uploads/sites/5/2016/12/Vol.-19-No.-1-Article-2-pp.6-9.pdf

Mitchell, C. M., Evanini, K., & Zechner, K. (2014). A trialogue-based spoken dialogue system for assessment of English language learners. In *Proceedings of the 5th international workshop on spoken dialog systems* (pp. 194–200). IWSDS. https://www.cise.ufl.edu/research/learndialogue/pdf/Mitchell-IWSDS-2014.pdf

So, Y., Zapata-Rivera, D., Cho, Y., Luce, C., & Battistini, L. (2015). Using trialogues to measure English language skills. *Journal of Educational Technology & Society*, *18*(2), 21–32. https://drive.google.com/file/d/1lL1TasxYQhqNIpLmXJH9IvVnxfs8OLtS/view

Song, Y., & Sparks, J. R. (2017). Measuring argumentation skills through a game-enhanced scenario-based assessment. *Journal of Educational Computing Research*, *56*(8), 1324–1344. https://doi.org/10.1177/0735633117740605

Yang, H. C., & Zapata-Rivera, D. (2010). Interlanguage pragmatics with a pedagogical agent: The request game. *Computer Assisted Language Learning*, *23*(5), 395–412. https://doi.org/10.1080/09588221.2010.520274

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: Volume 3. Authoring tools and expert modeling techniques* (pp. 169–178). U.S. Army Research Laboratory.

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. E. Boyer, M. Crosby, & K. Panourgia (Eds.), *Lecture notes in computer science: Intelligent tutoring systems, 12th international conference proceedings* (Vol. 6095, pp. 625–626). Springer. https://doi.org/10.1007/978-3-319-07221-0_84

Zapata-Rivera, D., Lehman, B., Jackson, G. T., & Liu, L. (2019, April 5–9). *Refining conversation-based assessments of science inquiry skills* [Paper presentation]. American Educational Research Association Annual Meeting, Toronto, ON, Canada.

Suggested citation:

Lopez, A. A., Guzman-Orth, D., Zapata-Rivera, D., Forsyth, C. M., & Luce, C. (2021). *Examining the accuracy of a conversation-based assessment in interpreting English learners' written responses* (Research Report No. RR-21-03). Educational Testing Service. https://doi.org/10.1002/ets2.12315

Action Editor: Donald Powers and Jonathan Schmidgall

Reviewers: Blair Lehman