Using Trialogues to Measure English Language Skills

Youngsoon So[1*], Diego Zapata-Rivera[2], Yeonsuk Cho[2], Christine Luce[2] and Laura Battistini[2]

[1]Department of English Language Education, Seoul National University, Seoul, South Korea // [2]Educational Testing Service, Princeton, New Jersey, USA // youngsoon_so@snu.ac.kr // dzapata@ets.org // ycho@ets.org // cluce@ets.org // lbattistini@ets.org

[*]Corresponding author

ABSTRACT

We explored the use of technology-assisted, trialogue-based tasks to measure the English language proficiency of students learning English as a second or foreign language. A presumed benefit of the system for language assessment is its suitability for use in scenario-based tasks that integrate multiple language skills. This integration allows test developers to simulate real-life language use situations by administering more authentic test tasks to strengthen the link between test performance and its interpretation as an indicator of expected success in performing real-life language use tasks. We conducted a usability study with 20 English language learners (ELLs) in grades 3 to 5, representing ten different first languages. Results suggested that the trialogue-based tasks engaged students in an authentic experience and provided evidence of their English proficiency. Perception data indicated that students perceived the trialogue-based tasks positively and understood how to respond to the virtual characters. Implications are discussed for future research and the potential of using trialogue-based assessment tasks for measuring the proficiency of young ELLs.

Keywords

Trialogues, Computer-assisted conversations, English language learners, English as a second or foreign language, Young children

Introduction

Conversations between a student and virtual characters have been used to facilitate learning by supporting human-like interactions with computer-based systems (Chan & Baskin, 1990; Graesser, Person, Harter, & Tutoring Research Group, 2001; Johnson, Rickel, & Lester, 2000). To date, this type of system has been used for non-assessment purposes, mainly for the tutoring of cognitive skills, such as scientific reasoning (Graesser et al., 2010). Our study examined the potential of applying such a conversation-based system to assess constructs that are different from those previously implemented for non-assessment purposes. More specifically, our study explored the use of conversation-based tasks as a means of gathering evidence of the English language proficiency of students learning English as a second or foreign language (ESL/EFL).

Trialogues, interactive conversations between one student and two virtual characters, have been used to create engaging, realistic environments in which a student is positioned to possess more or less information and/or knowledge relative to each of the two virtual characters. Graesser et al. (2010) utilized trialogues between a student and two virtual characters (a peer and a teacher) to identify evidence of student inquiry skills. In this particular setting, the student is expected to help the less knowledgeable virtual peer while receiving feedback or scaffolding from the more knowledgeable virtual teacher. Students can assume different roles within trialogue-based tasks depending on their expected prior knowledge, playing the roles of either the giver or the receiver of information (Cai et al., 2009).

We focus in this study on four potential advantages of applying the trialogue system to measure the English language proficiency (ELP) of young English language learners (ELLs). First, trialogue-based interactions take place in virtual environments. Such environments are familiar to today's young students, who are more immersed in new technologies than any previous generation (Kaiser Family Foundation, 2010). Well-designed virtual environments can involve young students in language-use situations that maximize meaningfulness, interactivity and engagement — qualities considered very important in teaching and assessing young language learners (Cameron, 2003; Hasselgren, 2000, 2005; McKay, 2005, 2006). Second, the trialogue system allows for the creation of virtual communication situations in which language learners must use language to perform the given tasks. Integration of obligatory communication situations into task design is not a new concept in second language teaching and learning. For example, information gap activities, which are commonly-used language teaching activities, deliberately

manipulate the information available to each student as a means to promote more genuine communicative interactions among language learners (Johnson, 1982; Widdowson, 1978). Third, the trialogue system can incorporate scaffolds or supports that help students demonstrate the best of their abilities within assessment tasks in a way that students perceive the scaffolds to be a natural part of their conversations with virtual characters.

Last, relative to traditional language assessment tasks, the trialogue system's virtual environment allows for the creation of authentic scenarios in which students use language to achieve communication goals that are similar to those which they are expected to achieve in non-testing situations. The last point is particularly important when considering the purpose of language testing, i.e., to obtain evidence of competence that can be extrapolated to non-test, real-life contexts (Bachman & Palmer, 2010). Therefore, assessment tasks that are similar to the language-use tasks expected to occur in non-test situations are more likely to lead to the valid interpretation and use of test results (Kane, 2006).

In the language assessment literature, it has been argued that integrated assessment tasks — tasks that require a student to engage multiple language skills, such as listening and speaking — are a more proximal simulation of non-test language use (e.g., Lewkowicz, 1997). To successfully participate in a conversation, one must perform the roles of both listener and speaker. However, current language assessments tend to measure listening, reading, speaking and writing in isolation, and the tasks in such assessments cannot easily implement language skill integration. This limitation is partially due to the lack of practical ways to create a test in which a student's response to a task can be used as input for the next task.

In light of the anticipated benefits of using trialogues to measure young ELLs' English language proficiency, the purpose of our project was to develop trialogue-based tasks in which 8- to 11-year-old ELLs could achieve communication goals while engaging multiple integrated language skills. We first report on the development process and describe the resulting tasks. Next, the results of a usability study conducted in March 2013 are discussed. We conclude with implications for language assessment and suggestions for future research.

Development process

The process used in the development of the trialogue-based tasks is summarized in Figure 1. The process began with identifying the language constructs of interest, followed by developing scenarios that required students to perform communication activities that engage the identified constructs. The developed scenarios were then implemented using a dialogue authoring tool. The implemented scenarios were combined with graphical components using the Unity system (Unity Technologies, 2013) to produce the tasks evaluated in the usability study. The development generally proceeded in the order presented in Figure 1, although considerations in a later stage made us revisit earlier stages. In the sections that follow, we show how each component of the development process was operationalized.

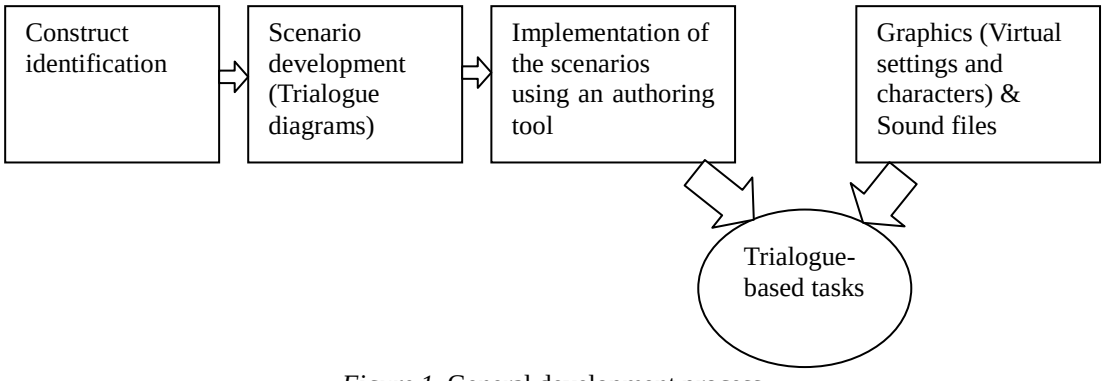

*Figure 1***.** General development process

Construct identification and scenario development

The target constructs were defined with reference to the learning objectives and communication goals commonly found in English curricula for elementary aged students in several countries where English is taught as a second or

foreign language, including Brazil, China, Korea, Japan, Mexico, and the Philippines. Table 1 summarizes the selected constructs by language skill.

Table 1. Constructs identified for the tasks

Language skill	Target constructs
Listening	• Understanding directions • Understanding simple questions
Speaking	• Answering simple questions • Identifying and summarizing key ideas • Forming an appropriate request
Reading	• Understanding written instructions

After selecting the constructs to be measured, the language use contexts in which students would be expected to interact with virtual characters while engaging the targeted language constructs were created. School settings (classroom and school library) were selected for the design of the language use contexts, and five trialogue-based tasks were developed, with each task designed to measure a combination of the constructs in Table 1. The tasks were designed such that different skills (e.g., understanding directions and answering simple questions) were integrated. Each task was designed to accommodate multiple possible conversation patterns between a student and the virtual characters, based on natural language processing of the student's response. All of these conversational possibilities are visually represented as "trialogue diagrams" (Figure 2).

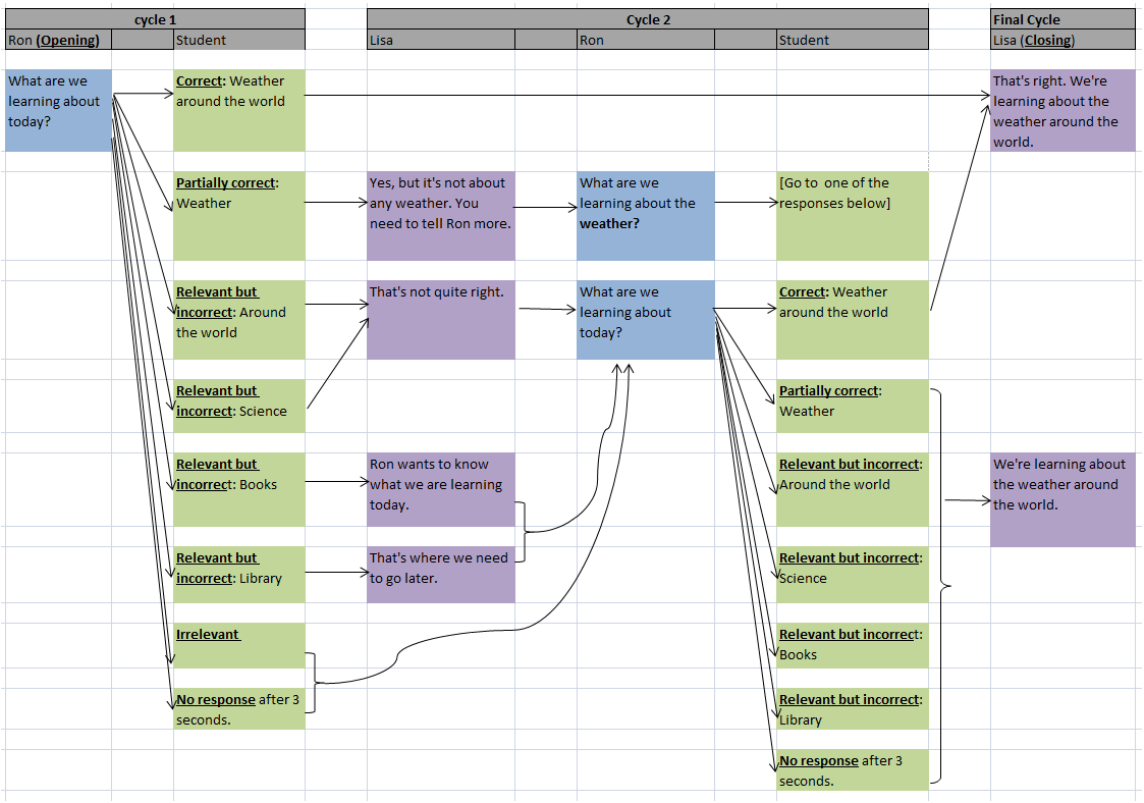

Figure 2. Trialogue diagram illustrating potential conversation flows

As illustrated in Figure 2, each task started with an opening that set the stage for interactions and ended with a closing that concluded the scenario and transitioned to the next task. The opening for each task was usually a question directed to the student. A range of possible student responses were considered and categorized as (a) correct, (b) partially correct, (c) relevant but incorrect, (d) irrelevant, or (e) no response. Each of these response categories was designed to trigger an appropriate next response from one of the virtual characters. For example, a correct response typically transitions to a closing statement, while a partially correct or incorrect response triggers a scaffold that provides the student with an opportunity to either elaborate on or correct the previous response. An irrelevant

response, or no response, triggers the repetition of the opening question. Each anticipated conversational pattern diagramed in Figure 2 was designed to provide evidence of the student's level of ability related to the target constructs.

In the specific task illustrated in Figure 2, three constructs were targeted: *understanding directions*, *understanding simple questions*, and *answering simple questions*. The first target construct was measured by examining the presence and amount of important information from the directions that were included in a student's response(s). For example, if a student provided a partially correct response in the first attempt but was unable to elaborate on the original response after a follow-up scaffold question was asked, that would be taken as evidence that the student did not fully understand the spoken directions provided and, therefore, possessed lower ability on the construct of *understanding directions*. The latter two constructs, *understanding simple questions* and *answering simple questions*, were measured by observing how well a student sustained the conversation with the virtual characters, as evidence of the student's understanding and appropriately responding to their questions. Thus, the fact that the student in the example above did not respond appropriately to the scaffold question would be taken as evidence that he or she did not understand the intention of the second question and/or did not know how to respond. This type of response pattern provided evidence that the student did not demonstrate high ability on the constructs of *understanding simple questions* and *answering simple questions*.

Scenario implementation and graphics integration

The scenarios developed in the earlier stage were implemented into computer-based conversations using a dialogue authoring tool based on AutoTutor Script Authoring Tool (Susarla, Adcock, van Eck, Moreno, & Graesser, 2003). Anticipated student responses were tested, and necessary revisions were made to ensure that the conversations between the student and the virtual characters proceeded as anticipated. The example in Table 2 illustrates one of the many conversation flows shown in Figure 2. The conversation starts with the virtual character Ron asking the student *What are we learning about today?*, to which the student provides the partially correct response *Weather*, which is then followed by scaffolds from the virtual characters. In this instance, the virtual character Lisa delivers the scaffold: *Yes, but it's not about any weather. You have to tell Ron more.* Lisa's instruction then triggers Ron to rephrase his original question: *What are we learning about the weather?* The student is then given the opportunity to elaborate upon the previous answer, saying *Weather around the world*. Finally, Lisa closes the conversation with an acknowledgement of the correctness of the response.

Table 2. Example of a trialogue conversation

Character	Utterance
Ron:	What are we learning about today?
Student:	Weather.
Lisa:	Yes, but it's not about any weather. You need to tell Ron more.
Ron:	What are we learning about the weather?
Student:	Weather around the world.
Lisa:	That's right. We're learning about the weather around the world.

The implemented conversations were integrated with graphics and sound files to produce the final tasks. The audio statements and questions produced by each virtual character were recorded using text-to-speech software and were incorporated into the tasks in audio format only; the written text was not displayed for the students. This decision was based on two related considerations. First, the real-life language use situations that the trialogue-based tasks are designed to approximate take place as oral communication. Second, measuring listening proficiency, in isolation from reading ability, is necessary to estimate a student's ability to participate in such conversations. By providing students with only the virtual characters' oral speech and excluding written text, a representation of the students' ability to perform non-test tasks can be inferred more accurately from their test task performance. Thus, written language was included in the tasks only when reading comprehension was part of the construct to be measured, as shown in Figure 3.

In the scenario shown in Figure 3, the student is expected to discuss one of the library rules with a virtual character (Ron) on the basis of understanding what is written on the board. A blinking arrow was added to the left side of the screen to draw the student's attention to the board. In this task, *understanding written instructions* (reading) is the

main construct measured, and two additional constructs, *understanding simple questions* (listening) and *answering simple questions* (speaking), are also measured through the conversation between a student and the virtual characters. It should be noted, however, that an automated speech recognition (ASR) engine had not yet been incorporated into the tasks at the time of the usability study reported in this paper. Because the technological capability to automatically recognize and accurately respond to students' spoken responses was not yet available, students were asked to provide two responses to each task: one spoken and one written. When students were directed by a virtual character to respond, they were first asked to record their spoken responses, and then to type what they said in a space that was provided. The spoken responses served two purposes: (a) to provide students with a more natural feeling of participating in an oral conversation and (b) to collect speech samples to train an ASR engine that would be used at a later stage of development in the project, discussed later in this paper. Thus, in the tasks used for data collection in 2013, the trialogue system used the students' written responses to determine the virtual character's next responses.

*Figure 3***.** Screenshot of a scene (Library scenario)

Evaluation: Usability study

The efforts to develop trialogue-based tasks were planned as a multiyear project with both long-term goals and short-term milestones. In the usability study, conducted in March 2013, we collected preliminary evidence on the usefulness of the trialogue-based tasks as a means to measure target constructs, with the following three objectives:

- to examine whether students tend to follow the "conversational flow" (see Figure 2) as expected, and if not, determine what modifications should be made to the current trialogue diagrams to better simulate how people actually speak in real life
- to gather lessons about the factors to be considered in developing additional trialogue-based tasks for the future
- to collect speech samples to be used to train an ASR engine for a future interactive speaking test

The first two objectives reflected short-term milestones, and the last objective was related to a more far-reaching, long-term goal to make the existing trialogue-based tasks into actual speaking tasks capable of supporting a spoken dialogue between a student and virtual characters. Therefore, the research questions of the study focused on collecting evidence related to the first two objectives:

RQ_1: Are the trialogue-based tasks usable for ELLs in grades 3 to 5? Can students in the target population interact with the trialogue-based tasks?
RQ_2: Can we gather useful evidence of particular English language constructs using trialogue-based tasks?
RQ_3: How engaging for students are the trialogue-based tasks compared to multiple-choice questions that target the same English language constructs?

Participants

Twenty students in grades 3 to 5, classified as ELLs by state criteria, were recruited from four public schools in a northeastern state in the United States. Student background information on gender, grade, first language, and English proficiency is summarized in Table 3. The proficiency levels were reported by an ESL teacher in each school; therefore, these teacher-reported proficiency levels may not be directly comparable across schools. However, they were useful as indicators of students' English language proficiency. The students represented a diverse range of first language backgrounds, with 10 first languages represented among the 20 participants. This criterion was considered critical in recruitment to ensure that the speech samples to be collected would have sufficient diversity to train an ASR engine at a later stage of the project.

Table 3. Background information of the participating students

Gender		First Language	
Female	8	Arabic	1
Male	12	Chinese	3
Grade		Danish	1
3rd	7	French	1
4th	8	Hindi	3
5th	5	Korean	2
English proficiency		Japanese	4
Beginning	6	Norwegian	1
Intermediate	11	Spanish	2
Advanced	3	Telugu	2

Materials

The following materials were used in the usability study.

Trialogue-based tasks: The tasks described in the *Development process* section of this paper were administered to each individual student.

Multiple-choice questions: Multiple-choice (MC) questions were chosen from an international standardized test that targets similar constructs to those targeted by the trialogue-based tasks: *understanding short spoken discourse*, *understanding short written text*, and *understanding short questions*. Ten listening comprehension questions and eight reading comprehension questions were selected. Although these questions are normally delivered in paper-and-pencil format for operational administration, the format was modified to be delivered via computer for the purposes of this study. The decision to perform this modification was made to control the variable of delivery medium between the trialogue-based tasks and the MC questions.

Background questionnaire: A background questionnaire was designed to gather participants' personal information (gender, age, and first language), English language learning experience, and computer use.

Usability questionnaire: Usability questions were used to investigate the clarity of the directions, the appropriateness of the time given to complete each task, the perceived difficulty of the tasks, the appropriateness of the speed of the listening input, and the overall experience of participants while they completed each task.

Engagement questions: Engagement questions investigated the participants' overall experiences with both types of tasks (i.e., trialogue-based and MC questions). The main focus was to examine whether participants perceived any differences between taking the trialogue-based tasks and the MC questions.

Procedure

Participating students were recruited with parental consent, and parents were asked to fill out the student background questionnaire on behalf of their children.

Over a two-week period, students participated in individual study sessions with two researchers per student, with one researcher acting as facilitator and administering the assessment tasks and the other acting as observer and taking notes on the students' behaviors and reactions. A typical study session lasted about 60 minutes and included the trialogue-based tasks, MC questions, and follow-up cognitive interviews using the usability and engagement questionnaires (see the *Materials* section). The order of the trialogue-based tasks and the MC questions was counterbalanced so that 10 participants were administered the trialogue-based tasks first and the other 10 students received the MC questions first. All study sessions took place at the students' schools and were recorded. Usability and engagement questions were administered immediately following each type of task, and engagement questions requiring comparison of the two types of tasks were administered at the end of the study session.

Results

Student responses and researcher observation notes were gathered during the 20 individual cognitive interview sessions. Data were analyzed to evaluate the evidence collected on the students' English language skills and task engagement. In this section, the results of the collected data are discussed in relation to each of the three research questions addressed in the study.

RQ_1: *Are the trialogue-based tasks usable for ELLs in grades 3 to 5? Can students in the target population interact with the trialogue-based tasks?*

This research question addressed the utility of the trialogue-based tasks for the intended student population by examining to what extent the participating students' familiarity with technology was related to their reactions to the trialogue-based tasks. In order to first investigate students' familiarity with technology, students were asked questions about their use of computers, either at home or at school, for four different purposes listed in Table 4. Results from these questions were later compared with students' reactions to the trialogue-based tasks to investigate any relationship. Table 4 summarizes the student responses.

*Table 4***.** Use of a computer for different purposes

	More than once a week	Once a week	Once a month	Never
Do homework	3	5	4	8
Play games	5	12	1	2
Use the Internet	11	5	3	1
Watch a movie, video, or DVD	9	3	3	5

With the exception of one student, an eight-year-old third grader who answered that he did not use a computer for any purpose (ID#19), the students all reported that they used a computer for at least one of the four purposes. Because the participants were familiar with computers to a certain extent, all of them were able to determine what they were expected to do to complete the trialogue-based tasks. Even when students did not immediately understand what to do, they eventually understood the task without any explicit assistance from the facilitator. All 20 participants completed the trialogue-based tasks. This fact provides evidence that the target student population is likely to be able to respond to the tasks used in the study without great difficulty in understanding how to complete the tasks.

RQ_2: *Can we gather useful evidence of particular English language constructs using trialogue-based tasks?*

This research question addressed whether the trialogue-based tasks gathered useful evidence of ability related to particular English language constructs. The question was examined from two perspectives. First, the students' performances on the trialogue-based tasks were compared with their levels of English language proficiency as identified either by their teachers or by their performances on the MC questions. Second, the same research question was indirectly addressed by investigating the extent to which the students engaged with the communication tasks as intended, with minimal influence by construct-irrelevant factors.

The collected evidence indicated that a positive association existed between students' English language proficiency and their performance on the trialogue-based tasks. First, three students who were identified as advanced ELLs by their teachers (ID#4, #10, and #14) performed very well on the trialogue-based tasks. They either answered all the questions correctly on their first attempt or were able to elaborate on their partially correct responses when the virtual characters provided scaffolds. On the other hand, six students who scored below average on the MC questions (ID#2, #3, #7, #8, #19, and #20) did not perform as highly on the trialogue-based tasks. For example, the student who scored the lowest on the MC questions (ID#2) could neither provide correct responses on the first attempt to any of the opening questions in the trialogue-based tasks nor correct his previous responses when given scaffolds. This tendency to not attempt to modify one's previous responses, even in the presence of scaffolded feedback, was found across the six students with lower scores on the MC questions. These two observations constitute preliminary evidence that trialogue-based tasks can provide evidence of ELLs' language proficiency.

In addition, responses to usability questions were analyzed to investigate whether students perceived trialogue-based tasks positively, which was assumed to be a necessary condition for the use of trialogue-based tasks in measuring English language constructs. Nine usability questions were asked at the end of the trialogue-based task session. Figure 4 summarizes the nine questions and the response frequencies.

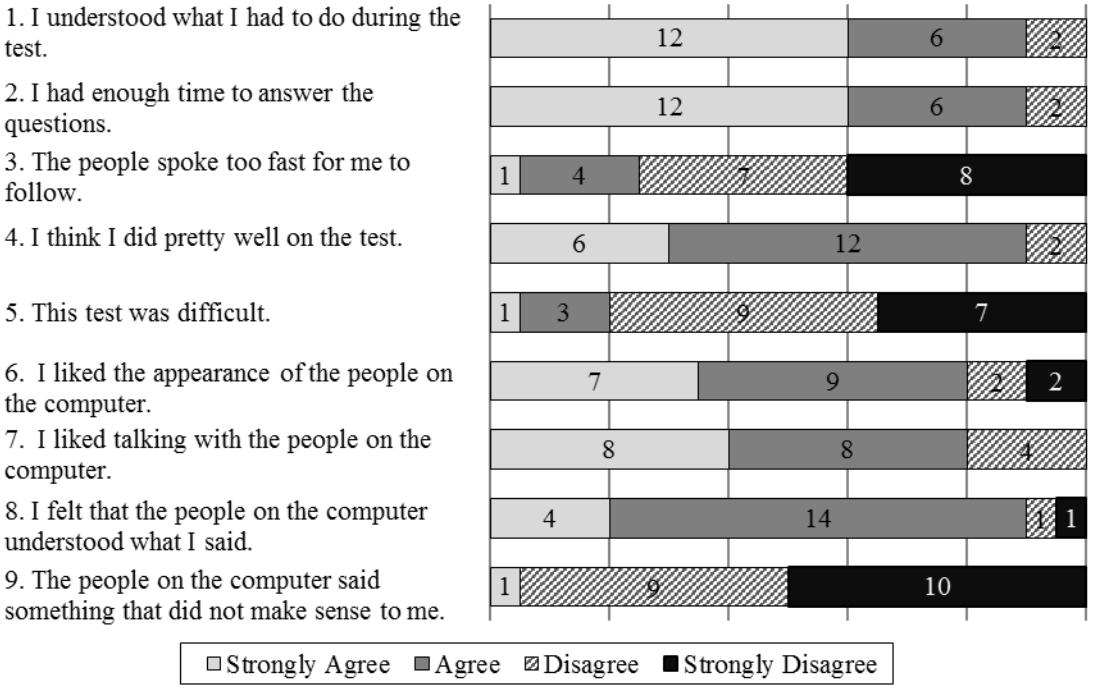

*Figure 4***.** Summary of responses to usability questions

The results generally suggested that students perceived the trialogue-based tasks positively. Participants reported that they understood how they were supposed to respond to the virtual characters (Q#1), and they had enough time to answer the questions (Q#2). Furthermore, students generally did not perceive the pace as too fast (Q#3) or think that the test was difficult (Q#5). Some particularly encouraging findings included many students reporting that they liked speaking with the virtual characters (Q#7) and that they felt that the virtual characters understood what they said (Q#8). Similarly, almost all students (19 out of 20) reported that what the characters said made sense (Q#9).

Despite these positive findings, the responses of the one student who reported not using a computer for any purpose (ID#19) were mostly negative. This student reported that he did not understand what to do in the trialogue-based tasks and that he did not feel that the virtual characters understood what he had said. This suggests that there may be a relationship between students' familiarity with computers and their experience taking the trialogue-based tasks.

During the usability study, students were given an opportunity to openly express their opinions about the trialogue-based tasks to capture any perceptions that were not elicited by the usability questions. Some students reported that they liked the trialogue-based tasks because they wanted to practice speaking and thought they could learn English better by doing this type of task. Two students (ID#10 and #13) reported that they liked the trialogue-based tasks because they simulated real life. However, other students shared negative feedback about some aspects of the trialogue-based tasks. For example, eight students reported that in some scenarios it was difficult to remember important information needed to answer the questions, and three students reported that they found the virtual characters' rate of speech to be too fast. In addition, three students did not like the fact that they had to speak (ID#1, #6, and #12). One student (ID#6) chose not to speak and instead only typed his responses. When asked, he reported that he did not want to speak. This finding may be attributable to the fact that some second language learners may not feel comfortable speaking in English.

Some areas for improvement emerging from the usability study will be considered in the later stages of this project in order to enhance the user experience. First, several students reported that they did not like the appearance of the virtual characters because the characters did not look realistic. Second, one student (ID#5) commented that he did not like the trialogue-based tasks because the characters kept asking him to repeat what he had already said. This comment suggests a possible limitation of the current system and/or of the conversation flows administered in the trialogue-based tasks. Despite such negative comments by a few students, the participating students were positive overall about the trialogue-based tasks. This result indicated that the trialogue-based tasks have the potential to be used as an innovative assessment tool to measure the English language proficiency of young ELLs.

RQ_3: How engaging for students are the trialogue-based tasks compared to multiple-choice questions that target the same English language constructs?

To address the final research question, a comparison was made of students' levels of engagement when using the trialogue-based tasks and the MC questions. Table 5 summarizes the students' responses to the three comparison questions.

Table 5. Summary of the responses to the comparison questions

Comparison question	Trialogue	MC questions
Which test did you find more difficult?	16	2
On which test do you think you did better?	2	16
Which test would you choose to take?	7	12

The results in Table 5 indicate that students reported thinking that they had performed better on the MC questions, and, if given the choice, they would choose to take the MC test over the trialogue test. This was a somewhat surprising finding given that student responses to the trialogue-based tasks were generally positive. Responses to follow-up probes revealed that the main reason students preferred the MC question format was due to their greater familiarity with that format and their consequent anticipation of better performance on the MC items.

Nevertheless, it should be noted that speaking was not required in response to the MC questions, and therefore, it may not be fair to compare student engagement between the two formats. This interpretation is strengthened by the fact that some students reported liking the MC questions because no speaking was required (ID#1, #6 and #12). These students' responses imply that second language learners may feel less confident using productive language skills (e.g., speaking) than receptive skills (listening and reading). However, there were students who preferred the trialogue-based tasks (ID#15, #16, and #18) despite the fact that their perceived level of performance was higher for the MC questions. These students reported preferring the trialogue-based tasks mainly because they felt that they could "practice more talking" and "learn English better," or they "wanted to be good at speaking." Another student who reported preferring the MC test for assessment (ID#13) explicitly stated that she would choose the trialogue-based tasks "if the purpose is learning, not testing." This student also commented that the trialogue-based tasks seemed to "simulate the real life [sic]." The students' perception that the trialogue-based tasks would help improve

skills that they personally wanted to build is discussed in the next section with regard to the implications of using such tasks in language assessment.

Discussion

The primary goal of this study was to investigate the usefulness of computer-mediated, conversation-based tasks, more specifically trialogue-based tasks, as a means of language assessment. The reported study provided initial evidence that students engaged with the interactions that were built into the trialogue-based tasks as anticipated, and that the students' experiences with the trialogue-based tasks were generally positive. In this section, implications of the results of the usability study are discussed from the following perspectives: (a) the representation of language use tasks in non-assessment situations, (b) potential construct-irrelevant factors, (c) the collection of formative information, and (d) the provision of meaningful communication opportunities for students with limited exposure to English communication.

Overall, the results of the usability study constitute evidence supporting the use of trialogue-based tasks with young ELLs as a means of providing a more engaging test-taking experience. The students' positive responses to the usability questions (see Figure 4) are important, because we assume that if students did not perceive the virtual characters as understanding and responding appropriately to their utterances, then the language samples elicited by the trialogue-based tasks would not bear much resemblance to non-test communicative language use. In addition, the finding that some students liked the trialogue-based tasks because the tasks "simulated real life" and, compared to MC questions, were "more realistic because the virtual characters talk" was very encouraging. When students perceive that assessment tasks are similar to the language use tasks that they perform in non-assessment contexts, it is more likely that they will engage in the abilities required for real-life situations. This finding is encouraging, as it hints at the potential of trialogue-based tasks to enhance the validity of interpretations made on the basis of student test performances. If, while completing trialogue-based language tasks, students engage in language and communication processes that are similar to those required in non-assessment situations, their assessment performances may be better indicators of their expected success in non-test, real-life communication tasks, thus providing a basis for more accurate and adequate inferences on the basis of test scores. Furthermore, the use of such tasks, being more proximal to real-life language use, may bring about a more positive washback effect in language learning and teaching by motivating language teachers to make greater use of tasks that help their students engage important language skills in more meaningful ways.

Despite these encouraging findings, several factors reported may have introduced construct-irrelevant variance in student performance on the trialogue-based tasks. For example, some students (ID#3, #6, #11, and #16) commented on the unrealistic appearance of the virtual characters. Although it is unknown to what extent the students' attention to the appearance of the virtual characters actually influenced their performance on the tasks, these comments suggest that the appearance of the virtual characters may have influenced student engagement during task interactions. Another potentially construct-irrelevant factor was related to memory load. Some students (ID#13 and #18) reported that they had difficulty retaining the information needed to successfully perform some of the tasks. It is important to note, however, that this situation mimics real-life experience and may actually contribute to the trialogue-based tasks' ability to capture the students' deployment of language skills in an authentic communicative interaction. It is not unexpected that such a task may place a somewhat greater burden on memory than the MC questions. Nonetheless, these findings provided useful insights into factors that should be taken into consideration when designing trialogue-based tasks to minimize the influence of potential construct-irrelevant factors and thereby maximize the validity of the interpretation of performances on the tasks as indicators of the target constructs.

With respect to the collection of formative information via the trialogue-based tasks, we would argue that the fact that scaffolds can be easily embedded within trialogue-based tasks makes the tasks potentially very useful for formative assessment. Depending on the specific conversational route that students take, they are given opportunities to correct or elaborate upon their previous responses by making use of differing amounts of the scaffolding provided. One student (ID#4) reported that he liked the trialogue-based tasks because "the characters respond [to] a wrong response." This example clearly illustrates that the student was able to notice, on the basis of the characters' reactions to his response, that something was not quite right. Furthermore, it was observed that not every participating student responded to the same scaffolds in the same way, and not all students were able to utilize the additional scaffolds provided. This variability implies that student ability levels on the target constructs can be

differentiated depending on the specific ways students interact with trialogue-based tasks (i.e., whether they provide a correct response in the first attempt and whether they benefit from the provision of additional scaffolds). This design feature is important because it allows us to gather more information on where a student's strengths and weaknesses lie, indicating areas on which students should focus for improvement.

Finally, we are confident that the communicative tasks investigated in this study hold potential to provide meaningful English learning opportunities to students learning English in countries where exposure to English language is generally restricted to classes at school. These English classes are often taught by non-native English speaking teachers who share a first language with their students. Students in such contexts may encounter relatively limited opportunities in which they are required to use English. This point is well illustrated by comments from students (ID#15, #16, and #18), who said that they found the trialogue-based tasks useful to practice speaking in English. This finding, that even students who are learning English in an English-speaking country seek more opportunities to practice speaking, provides a strong case for the need for such learning opportunities for students learning English in countries where English is not the primary or official language. Related to this point, the data collected from the usability study showed that some students were reluctant to speak and preferred the MC questions because they did not require any spoken responses (ID#1, #6, and #12). This finding indicates that, through trialogue-based tasks, students can gain opportunities to practice the skills in which they have less confidence and need more practice. We expect such opportunities to be helpful in students' goal of developing well-balanced language proficiency across all language skills. We also note that, in some instances, the trialogue-based tasks may hold an advantage over face-to-face communication for students who feel less intimidated when interacting with virtual characters than with real humans. This last point is particularly relevant for young students, who are the target population of the present project, as achieving positive affective reactions with this age group is thought to be essential for gleaning meaningful performances from assessment tasks (McKay, 2006).

This study had several limitations. No inferential statistical analysis was conducted to investigate the relationship of student performances on the trialogue-based tasks with other criteria of their English proficiency. The decision not to perform the statistical analysis was made on the basis of the following three reasons. First, the sample size (N = 20) was small. Second, the participating students performed very well on the MC questions (*Mean* = 17.0 out of 18 possible points, SD = 1.39). Given this narrow distribution of scores on the MC questions, meaningful results from inferential statistics were not expected. The third reason for not conducting a statistical analysis was that the methods used to score responses to the trialogue-based tasks need further validation. This is due to the fact that the trialogue-based tasks are very different from traditionally used language assessment tasks, and more research is needed regarding appropriate methods to score their responses. We discuss this point further in the next section, which addresses the future direction of the research and development project.

Conclusions and future work

Trialogue-based tasks have the potential to be used as tools to measure the interactive nature of language use, as discussed in the present paper using the data collected from a usability study. The types of tasks designed for the trialogue system require not only that students contribute to the tasks, but also that they receive immediate feedback on their performances. The performance feedback provided by the virtual characters can also serve as scaffolding, allowing the students to demonstrate their abilities both with and without this additional support.

Planned future work includes incorporating an ASR system to accept spoken input directly, freeing students from the need to type their responses. If an ASR system can be implemented, the trialogue-based tasks are expected to function in a more engaging and authentic manner, enhancing their usefulness as a measure of speaking ability. Subsequent to the data collection reported in this paper, an ASR system has been developed for the tasks, and a usability study was conducted with the ASR-enabled version. The results from the ASR-enabled version are being analyzed at the time of the writing of this paper, and they are expected to be disseminated in a future publication. In addition, scoring models are currently being validated to score responses collected through trialogue-based tasks. Given the more integrated nature of such tasks, scoring methods that have been used for more traditional discrete tasks may not be readily applicable to responses collected from the trialogue-based tasks. Next, additional trialogue-based tasks are also under development to measure language constructs that were not targeted in the tasks discussed in this paper. Finally, a larger-scale data collection is scheduled with an expanded version of the assessment that is ASR-enabled and includes more tasks targeting the measurement of more constructs.

Acknowledgements

The research reported in this article was conducted when the first and corresponding author was employed by Educational Testing Service. The authors would like to thank the editor and the two anonymous reviewers of *Educational Technology & Society* for their feedback. The authors would also like to acknowledge their external collaborators Art Graesser, Carol Forsyth, and Zhiqiang Cai from The University of Memphis, as well as Kyle Staves and his colleagues from SortaSoft, LLC, for their assistance in the instrument development. Finally, the authors acknowledge Don Powers, Irvin Katz, Alexis Lopez, Tanner Jackson, Kristin Williamson Worden, and Ian Blood for their useful comments on an earlier version of this paper.

References

Bachman, L., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford, UK: Oxford University Press.

Cai, Z., Graesser, A. C., Millis, K. K., Halpern, D., Wallace, P., & Moldovan, C. (2009). ARIES!: An intelligent tutoring system assisted by conversational agents. In V. Dimitrova, R. Mizoguchi, B. DuBoulay & A. C. Graesser (Eds.), *Proceedings of 14th International Conference on Artificial Intelligence in Education: Building learning systems that care: From knowledge representation to affective modelling* (p. 796). Amsterdam, the Netherlands: IOS Press.

Cameron, L. (2003). Challenges for ELT from the expansion in teaching children. *ELT Journal, 57*(2), 105–112.

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., Britt, A., Millis, K., Wallace, P., Halpern, D., Cai, Z., Kopp, K., & Forsyth, C. (2010). Critiquing media reports with flawed scientific findings: Operation ARIES!, a game with animated agents and natural language trialogues. In J. Aleven, J. Kay & J. Mostow (Eds.), *Lecture notes in computer science* (pp. 327–329). London, UK: Springer.

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.

Hasselgren, A. (2000). The assessment of the English ability of young learners in Norwegian schools: An innovative approach. *Language Testing*, *17*(2), 261–277.

Hasselgren, A. (2005). Assessing the language of young learners. *Language Testing, 22*(3), 337–354.

Johnson, K. (1982). *Communicative syllabus design and methodology*. Oxford, UK: Pergamon.

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.

Kaiser Family Foundation (2010). Generation M2: Media in the lives of 8- to 18-year-olds. Retrieved March 25, 2014, from http://www.kff.org/entmedia/mh012010pkg.cfm

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education & Praeger.

Lewkowicz, J. A. (1997). The integrated testing of a second language. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education, vol. 7: Language testing and assessment* (pp. 121–130). Dortrecht, the Netherlands: Kluwer.

McKay, P. (2005). Research into the assessment of school-age language learners. *Annual Review of Applied Linguistics*, *25*, 243–263.

McKay, P. (2006). *Assessing young language learners*. Cambridge, UK: Cambridge University Press.

Susarla, S., Adcock, A., van Eck, R., Moreno, K., & Graesser, A. C. (2003). Development and evaluation of a lesson authoring tool for AutoTutor. In V. Aleven, U. Hoppe, J. Kay, R. Mizoguchi, H. Pain, F. Verdejo & K. Yacef (Eds.), *AIED 2003 Supplemental Proceedings* (pp. 378–387). Sidney, Australia: University of Sydney School of Information Technologies.

Unity Technologies (2013). *Unity*, Version 3.5.7f6 [Computer Software]. San Francisco, CA: Unity Technologies.

Widdowson, H. G. (1978). *Teaching language as communication*. Oxford, UK: Oxford University Press.