Conversation-based Assessments: An Innovative Approach to Measure Scientific Reasoning

Lei Liu, Jonathan Steinberg, Farah Qureshi, Isaac Bejar, and Fred Yan

***Abstract*— This study explored the scalability of an innovative assessment approach, conversation-based assessment, to measure constructs related to scientific reasoning, where virtual students interact with students to demonstrate their abilities to use evidence to support a prediction. We designed two parallel tasks using the same design pattern and compared the equivalence of these tasks. We designed an experimental study to investigate whether the two parallel tasks performed equivalently by comparing student performance on both tasks. The results of the study shed light on the challenges of scalability of conversation-based assessment and the implications for next steps.**

***Index Terms*—Conversation-based assessment, virtual agents, scientific reasoning**

I. Introduction

NEW views of learning (e.g., constructivism and activity theory) call for innovative assessment modes that track students' thinking processes more closely. One underlying assumption of these new learning theories is that knowledge is socially constructed and mediated through communicating ideas with others. Conversation-based assessment is one promising approach to track and record students' higher-order thinking processes such as scientific reasoning. Building on previous research on natural language intelligent tutoring systems (Graesser, Person, & Harter, 2001), a conversation-based approach entails computer-mediated conversations. Students may take different paths through scripted conversations based on the content of their responses. The present study focused on the scalability of a conversation-based assessment approach by posing the following specific question: how can the production of conversation-based assessments be scaled up to measure higher-order thinking across different scientific topics? This problem is important given the cost-effective and validity concerns of innovative assessments, which is an area that lacks research. We describe an approach evaluating scalability whereby a successful implementation is used as the basis for developing parallel or isomorphic tasks, an approach that has been found useful in domains as diverse as architecture (Bejar, 2002) and networking (Kunze, Mehta, & Levy, 2015). In our study, we focused on within-domain scalability, namely, building a parallel task (Weather) to our base conversation-based task in the science domain measuring the same set of constructs (Volcano), and evaluating the psychometric equivalence of the tasks. We designed an experimental study to investigate whether these two science tasks performed equivalently and therefore could be used interchangeably. Specifically, we conducted preliminary and confirmatory analyses of the data. The preliminary analysis consisted of item analysis and exploratory factor analysis; the confirmatory analysis evaluated whether performance, response time, and internal structure were comparable across the two versions of the task.

Manuscript received May 25, 2016.

L. Liu is with ETS, Princeton, NJ 08541 USA (609-734-5183; e-mail: lliu001@ets.org).

I. Bejar is with ETS, Princeton, NJ 08541 USA. (e-mail: ibejar@ets.org).

F. Qureshi is with ETS, Princeton, NJ 08541 USA. (e-mail: fqureshi@ets.org).

J. Steinberg is with ETS, Princeton, NJ 08541 USA. (e-mail: jsteinberg@ets.org).

F. Yan is with ETS, Princeton, NJ 08541 USA. (e-mail: fyan@ets.org).

II. Conversation-based Assessment

A. Trialogues

Conversations have been used in assessment mostly in language proficiency exams (Wong, 2000). However, it is possible that these can also be used to measure student cognitive skills in different domains. Conversations have been frequently used to support formative assessment practices in instruction for collecting evidence of students' understanding by tracking and recording students' higher-order thinking process over time. In computer-supported communication (e.g., chat box format), displaying the conversation history allows students to access their cognitive processes when communicating with virtual agents or real partners. Given all the above opportunities, conversations can be applied in assessments to support eliciting evidence of student cognition. In our study, we applied one kind of conversation-based assessment – trialogues – in our assessment design.

A trialogue is a form of adaptive testing where the student is assessed by means of one or more simulated conversations with two virtual agents. Underneath the conversation is a tree-like structure or conversation-space diagram (Zapata-Rivera, Jackson, & Katz, 2015) that contains all the possible interactions designed to elicit student knowledge on certain topics. For example, students can be assessed on their scientific thinking skills such as using evidence to support and argue for a prediction, through conversing with virtual agents in the trialogue assessment. Based on students' response to a

question that asks them to make a prediction, students are branched through the conversation space to elicit their evidence base for that prediction. Students also get a chance to revise their original prediction during their reasoning process. Some characteristic features of conversation-based item types, such as providing immediate feedback through virtual agents, can lead to greater student motivation. This is a desired goal for assessments because more accurate measurement is derived from students "trying their best", and possibly leading to more valid student assessment. However, in practice, such benefits may not be affordable unless the tasks can be developed economically, and more importantly proven as effective measures of higher-order thinking which is often hard to measure through traditional measurement. To explore the scalability of the innovative assessment approach, we developed two parallel prototypes with the goal of evaluating their equivalence in assessing student knowledge, skills, and abilities with respect to both Earth Science knowledge and related inquiry skills.

B. Parallel trialogue tasks

The development of the base trialogue Volcano task was informed by a design pattern and designed based on evidence-centered design principles. The purpose of the parallel prototype development was to explore whether conversation-based assessment is a scalable and valid approach to measure higher-order thinking. Design patterns are tools that can help task designers to think through substantive aspects of an assessment argument and can help capture design rationales in a re-usable and generative form (Mislevy & Haertel, 2006). The design pattern aimed to support the construction of computer-delivered conversation-based tasks that engage students in conversational interactions with virtual agents and data collection tools. The design pattern focused on the skills of planning and carrying out data collection in the virtual field and using collected data as evidence to predict a natural event. The Volcano task asked students to design and carry out a data collection plan for the purpose of making a prediction of the likelihood of a volcanic eruption. When developing the parallel Weather task, we used the same design pattern to guide our prototype development about collecting data for making a prediction of the likelihood of a thunderstorm.

Our design pattern focuses on a subset of components of scientific reasoning, in particular using observation data to make a prediction for natural events, and includes four focal areas of knowledge, skills, and abilities (KSAs) relevant to scientific reasoning: earth science knowledge, analyzing data and identifying patterns, conducting data collection, and making predictions based on data. All items in both prototypes were mapped onto these four constructs that are aligned with the new science standards.

III. STUDY DESIGN

To evaluate the equivalence of the Weather and Volcano prototypes in measuring the same constructs, we administered both Volcano and Weather tasks to 210 students in the fall of 2014. Students were randomly assigned to one of two versions of alternating order of the same assessment. Students with an odd-numbered student ID were in one condition and students with an even numbered student ID were in the other condition. Below are three specific research questions to explore the equivalence of the two tasks:

1. Do the two prototypes meet basic psychometric criteria?
2. How do the relationships among constructs compare across the two prototypes?
3. Do randomly equivalent samples of students perform comparably on both trialogue tasks with respect to the focal constructs and the time demands of the task?

A 2x1 between-condition design was applied in the study. Students were randomly assigned to two conditions. In condition 1, students took the Volcano task first and then took the Weather task. In condition 2, the order of the tasks were counterbalanced. All participants took the background information questionnaire (BIQ) at the beginning of the study and a post-survey at the end of the study. Both instruments were administered online. The BIQ survey consisted of demographics, self-reported course grades, grade level, technology use, prior game use, prior knowledge of volcanoes and weather, and non-cognitive attributes (e.g., persistence, willingness to learn, domain relevant self-efficacies, interest in domain, etc.). The post-survey included a combination of constructed response items and multiple choice items that measured both conceptual understanding and science inquiry skills related to the constructs of the Weather and Volcano prototypes. In addition, the posttest also included questions to collect evidence of participants' perceived motivation and engagement when interacting with trialogue systems.

IV. DATA ANALYSIS

A. Basic psychometric criteria: Item and Exploratory Factor Analyses

We first conducted basic analyses of the psychometric functioning of the two versions of the task. Specifically, we conducted item analyses (Penfield, 2013) and exploratory factor analyses. Item analysis evaluates the difficulty of each item as well as its discrimination power (i.e., is the item able to sort students into those that know the answer to the item and those that do not). The results of the item analysis are relevant to the scoring assumption in a validity argument (Kane, 2006) and provides an aspect of the foundation for a validity argument regarding the scores from the task. In addition, given that the items constructed for each prototype were intended to be equivalent, and administered to randomly equivalent samples, we explored whether the item difficulties and discrimination were comparable. We followed the item analysis with an exploratory factor analysis of the inter-relationships among items to determine the structural similarity of the Volcano and Weather tasks.

B. Relationship among constructs: Confirmatory analyses

We evaluated comparability in the level of student performance exhibited on the Volcano and Weather tasks. As the data for the initial analyses consisted of observed scored responses on multiple-choice items as either right (1), wrong

(0), or in some cases partial credit (0.5 or 0-3 in the case of constructed response items), exploratory factor analyses (EFA) were conducted separately on each form in SAS 9.3 using a polychoric correlation matrix first generated from PRELIS 2.57. Within-prototype confirmatory factor analyses (CFA) were conducted in LISREL 8.80 testing for the fit of the data to a single latent dimension.

C. Performance analysis: Task scores and response times

For purposes of comparing the equivalence of Volcano and Weather prototypes, we conducted descriptive analyses to compare the mean construct scores as well as the total scores across the two prototypes. Construct sub-scores were computed by simply summing the individual item scores and then a normalized mean was computed (sum/items), also regarded as percent correct. We also compared the total time spent on each of the two prototypes.

V. RESULTS

A. Preliminary Item and Factor Analyses

The item analysis examined the difficulty and discrimination of the scored items in each task. The ranges of difficulty were 0.06-0.98 for Volcano and 0.05-0.81 for Weather. The ranges of discrimination were 0.34-0.84 for Volcano which was good and 0.08-0.88 for Weather which indicated some underperforming items. There was also a difference in presentation for one item between tasks: Weather is a single selection versus the counterpart in Volcano which is a multiple selection item, making the item inherently more difficult in Volcano, possibly due to a higher tendency to misread or misunderstand the question (Cassells & Johnstone, 1984; Haladyna & Downing, 2002).

Results from the exploratory factor analysis for each prototype indicated the presence of a relatively strong first factor but more emphasis was placed on the confirmatory factor analysis to allow items to load on four particular constructs: Earth Science Knowledge, Analyzing Data & Identifying Patterns, Conducted Data Collection, and Making Predictions Based on Data. A fifth construct, Evidence-Based Reasoning while, part of the assessment, was not included for this analysis, but is discussed with respect to the performance analysis.

B. Confirmatory Factor Analyses

The fit statistics from the within-prototype confirmatory factor analyses (CFA) shown in Table 1 were all better for Weather than Volcano.

Table 1. CFA Fit Statistics for Volcano and Weather.

	Volcano	Weather
RMSEA (90% CI)	0.085 (0.070, 0.101)	0.074 (0.058, 0.090)
CFI	0.880	0.949
SRMR	0.145	0.133

Note: RMSEA = Root Mean Squared Error of Approximation; CFI = Comparative Fit Index; SRMR = Standardized Root Mean Residual.

The latent inter-correlations shown below in Table 2 indicate low to moderate relationships for each prototype, suggesting four constructs may exist, even though the magnitudes may differ

Table 2. Inter-correlations among postulated skills for Volcano and Weather.

	Volcano				Weather			
Factor	ESK	ADP	CDC	MP	ESK	ADP	CDC	MP
ESK	1.000	- -	- -	- -	1.000	- -	- -	- -
ADP	0.432	1.000	- -	- -	0.802	1.000	- -	- -
CDC	0.223	0.422	1.000	- -	0.545	0.580	1.000	- -
MP	0.319	0.115	0.232	1.000	0.397	0.249	0.384	1.000

Note: ESK = Earth Science Knowledge**;** ADP = Analyzing Data & Identifying Patterns; CDC = Conducting Data Collection; MP = Making Predictions Based on Data.

The efficacy of the task model with multiple dimensions appears to be relatively good. For Volcano, all but one item loads above 0.40 on these latent dimensions, while for Weather all items load above 0.33. Additionally, there is evidence that the items generally load saliently on individual prototype factors, but the model fit is better with four factors for each prototype. Therefore, while one can examine the prototype as a whole, its underpinnings as represented by these four constructs are worthy of examination too.

C. Performance Analysis

As shown in Figures 1 and 2 using a percent correct metric, in general and regardless of task order, students performed better in Weather (54-56%) than on Volcano (47-49%) and particularly in Earth Science Knowledge (Weather = 41-50%; Volcano = 27-30%), possibly because that students had more opportunities to learn about weather knowledge in their daily lives. Aside from that, students' performance on items related to some constructs (e.g., Conducted Data Collection and Making Predictions Based on Data) was similar across prototypes, but performance on other items was not comparable related to Analyzing Data & Identifying Patterns (Weather = 79-84%; Volcano = 68-73%) and Evidence-Based Reasoning (Weather = 41-62%; Volcano = 12-26%), again partly because those constructs have more demands on applying Earth Science Knowledge.

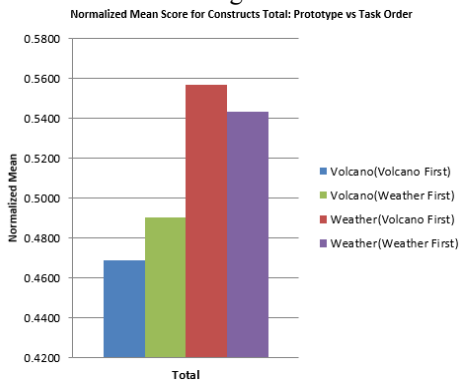

Fig. 1. Normalized mean scores for total scores across prototypes in two conditions.

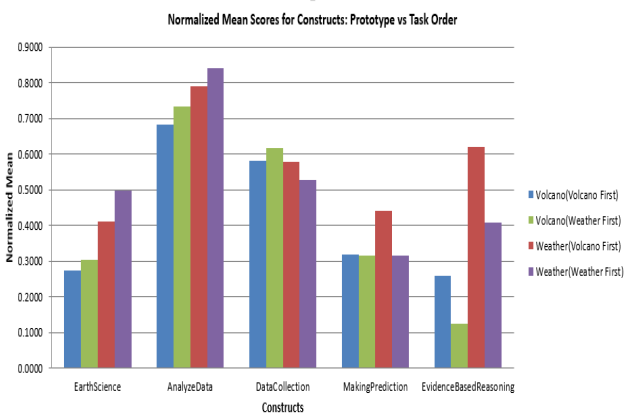

Fig. 2. Normalized mean scores for constructs across prototypes in two conditions.

In general, students spent more time (minutes) on Volcano (Mean = 30.2; SD = 13.4) than on Weather (Mean = 27.5; SD = 11.8), which can also be explained by a possible difference in familiarity with the contexts. However, the correlation in time spent between prototypes was close to zero. This necessitated the analyses to be broken out by task order. Time (minutes) spent on the second prototype was shorter (Mean = 23.6; SD = 7.1) than for the first prototype (Mean = 34.1; SD = 14.8) with a correlation of 0.32. The correlation in task times when Volcano was administered first was 0.35 and when Weather was administered first, it was 0.28.

VI. CONCLUSION

Even with all the limitations of the study (e.g., a limited data set), as shown in both item analyses and factor analyses, there are some encouraging results to show evidence of the comparability across two science conversation-based prototypes that were designed to measure similar constructs in two different science contexts. At the item level, both the item discrimination and difficulty were very much similar across the two prototypes. Performance on the aggregation of items into construct sum-scores also suggested a high degree of comparability. The analysis of the item-level performance through factor analysis suggested a high degree of cohesion among the items reflected in the strong presence of a primary dominant factor. In addition, there was evidence that students' performance on items related to some constructs (e.g., data collection and making predictions) was comparable, but performance on items related to other constructs (e.g., earth science knowledge, analyzing data, evidence-based reasoning) was not comparable. Our interpretation is that those incomparable items share highly similar surface level features (e.g., multiple choices) which may have enhanced student performance despite isomorphic item design. However, the comparable items mostly include conversation-based and performance-based items, indicating the potential advantage of using such items to measure student understanding in different contexts (Zapata-Rivera et al., 2015).

Furthermore, the distribution of response times was reasonably comparable across the two prototypes. In both conditions, students spent less time on the second prototype, suggesting some increased familiarity with the interface. In general, students performed better on the Weather prototype than the Volcano prototype, possibly because students were more familiar with the content through their daily life experiences. Such departures from strict isomorphism are to be expected and do not necessarily preclude the possibility of using the two as if there were exchangeable, especially in a low-stakes assessment environment. The preliminary results showed promise with respect to scalability in both developing parallel prototypes in the domain of science. The possible use of similar conversation-based structures to develop assessments in different contexts indicate possible cost savings for innovative assessment development. We also realize the limitation of our research given the small sample size. For future studies with a larger sample size, the underlying issues encountered in the factor analyses with the use of polychoric correlations and other analyses could be resolved through MIRT procedures.

REFERENCES

[1] Bejar, I.I. (2002). Generative testing: From conception to implementation. In S. H. Irvine & P. C. Kyllonen (Eds.), Item generation for test development (pp. 199-218). Mahwah, NJ: Earlbaum.

[2] Cassels, J. T., & Johnstone, A. H. (1984). The effect of language on student performance on multiple-choice tests in chemistry. Journal of Chemical Education, 61, 613-615.

[3] Graesser, A. C., Person, N. K., Harter, D. (2001) The Tutoring Research Group: Teaching tactics and dialogue in AutoTutor. International. Journal of Artificial Intelligence in Education 12, 257-279.

[4] Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiplechoice item-writing guidelines for classroom assessment. Applied Measurement in Education, 15, 309-334.

[5] Kane, M.T. (2006). Validation. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 17-64). Westport, CT: Praeger Publishers.

[6] Kunze, K. L., Mehta, V., & Levy, R. (2015). *Leveraging psychometric isomorphism in assessment development.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL, April 2015.

[7] Mislevy, R. J., & Haertel, G. D. (2006). Implications of Evidence-Centered Design for Educational Testing. Educational Measurement: Issues and Practice, 25, 6-20.

[8] Penfield, R.D. (2013). Item analysis. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise & M. C. Rodriguez (Eds.), APA handbook of testing and assessment in psychology, Vol. 1: Test theory and testing and assessment in industrial and organizational psychology (pp. 121-138). Washington, DC: American Psychological Association.

[9] Wong, J. (2000). Delayed next turn repair initiation in native/non-native speaker English conversation. Applied Linguistics, 21(2), 244-267.

[10] Zapata-Rivera, D., Jackson, T., & Katz, I.R. (2015). Authoring conversation-based assessment scenarios. In R. A. Sottilare, A. C. Graesser & J. Hu (Eds.), Design Recommendations for Intelligent Tutoring Systems (pp. 169-178). Army Research Institute.