Linguistic Profiles of Students Interacting with Conversation-Based Assessment Systems

Carol M. Forsyth, Jesse R. Sparks, Jonathan Steinberg & Laura McCulla
Educational Testing Service
[cforsyth,jsparks,jsteinberg,lmcculla]@ets.org

ABSTRACT
Conversation-based assessment systems allow for students to display evidence of their knowledge during natural language conversations with artificial agents. In the current study, 235 middle-school students from diverse backgrounds interacted with a conversation-based assessment system designed to measure scientific inquiry. There were two versions of the conversations where the initial question was manipulated to examine the relationship between question-framing and conversational discourse. We analyzed the human input during these conversations post-hoc using LIWC to discover linguistic profiles of students that may be related to the type of question asked as well as overall task performance. Furthermore, we compared these linguistic profiles to human ratings as a validity check and to inform our interpretation. Results indicated four separate profiles determined by linguistic features that indeed align to human scores and performance in directions consistent with the effects of question framing. These results offer important implications for improved detection of types of student learners based on linguistic features that do not differ by diverse student characteristics and for designing conversation-based assessments.

Keywords
Conversation-based Assessment, Computational Linguistics, Artificial Agents, Question-Framing

1. INTRODUCTION

1.1 Conversation-Based Assessment
Conversation-based assessments (CBAs) are interactive formative assessment systems with natural language conversations between humans and two or more artificial agents designed to capture evidence of a student's knowledge, skills, and abilities (KSAs) [36]. CBAs leverage previous research on digital learning environments with similar artificial agents or "talking heads" [5, 10, 19, 20, 23,35], artificial intelligence and technology enhanced assessments [4, 6, 26] to create environments where students take actions, answer questions, and participate in conversations to display their KSAs. CBAs include other components of simulated, scenario-based, and game-based environments [30, 31, 35] in addition to conversations.

Learning environments with natural language conversations have aided increasing student motivation and deep learning as students converse with artificial agents [1, 2, 21]. Several types of adaptive conversations have been created to accomplish this goal [17, 18] but we will focus on AutoTutor [13] as it greatly influenced the creation of CBAs. AutoTutor is an Intelligent Tutoring System where students have natural language tutorial conversations with an artificial agent; this system shows learning gains comparable to expert tutors in dozens of experiments across multiple domains [13]. Perhaps the "secret sauce" is the adaptive scaffolding moves, which are based on extensive analysis of expert tutor and student interactions [13,15] which include providing hints or broad clues, prompts requiring single word or phrase answers, and assertions as part of the scaffolding framework with associated natural language processing (NLP), which is beyond this paper to discuss [see 13].

CBAs augment the original conversational framework while utilizing the associated NLP of AutoTutor to create more constrained conversations for gathering evidence of KSAs which is necessary for assessment, where less information can be given to the student than in a tutorial dialogue. CBA conversations are designed to elicit information from students that may be difficult to elicit via other means such as multiple-choice or open-ended items. An important issue for CBA design is the impact of question framing in eliciting information from students.

1.2 Question Framing
The need to capture evidence of student knowledge has spurred research in question-asking for decades [3, 8, 15, 22]. This research has yielded multiple taxonomies of questions that seek to elicit student knowledge with respect to both mental representations and cognitive processes [12, 14, 27, 28, 29]. For example, Bloom's taxonomy [3] is well-known for capturing depth of understanding [7]. Specifically, multiple-choice questions provide evidence of shallow or factual understanding while open-ended questions require a deeper conceptual understanding to provide sufficient answers.

Formulating main questions to initiate a conversation with an artificial agent that may require multiple turns to elicit evidence from students is quite different from formulating a single question asked with a constructed response item. Therefore, we drew from a detailed taxonomy [15] to consider specific types of questions that may help elicit information and inform cognitive processing [14,

C. Forsyth, J. Sparks, J. Steinberg, and L. McCulla. Linguistic profiles of students interacting with conversation-based assessment systems. In A. Mitrovic and N. Bosch, editors, *Proceedings of the 15th International Conference on Educational Data Mining*, pages 594–600, Durham, United Kingdom, July 2022. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.6852980

15]. This taxonomy includes 16 question types at varying levels of depth. Within a CBA task, however, constraints including the target constructs, context, and scenario narrow the possibilities of questions that are appropriate for assessment (e.g., comparison and justification questions). Notably, research shows that tasks requiring students to generate justifications for their responses, such as constructing arguments, can sometimes lead to better learning than comprehension tasks [11]. Therefore, we investigated student responses to two separate types of question framing in a CBA task for science inquiry that prompted students to make a comparison between two artifacts (i.e., observation notes collected from simulated data), contrasting an approach where students were asked to *make a selection and explain their choice* (comparison-framing) with one in which students are asked to *justify whether or not they agree with a choice made by a virtual agent* (argumentation-framing) [34]. Previous research suggests that this manipulation had an impact on CBA conversation performance with the argumentation-framing condition making better selections but showing poorer explanations than the comparison-framing condition [33].

Previous research on CBAs suggests that additional computational linguistic analysis beyond the NLP algorithms needed to operate the system can provide fine-grained insights on how students respond [9]. Furthermore, CBAs offer an opportunity to explore methods for inferring information about students' experiences from their responses in ways that may have implications for equity issues in assessment, given the wide range of responses that can be accepted and awarded credit. Thus, in this study we analyzed responses from a diverse sample with a bottom-up approach using linguistic features and associated profiles, contextualized them within question framing conditions, and compared these profiles based on human scoring and on task performance.

2. METHODS

2.1 Participants

In total, 235 middle school students were recruited from one rural (35%) and one urban (65%) school, yielding quite a diverse sample. Specifically, the sample was 48% female and 52% male, 34% free/reduced price lunch eligible, and 79% White, 9% Black/African American, 4% Hispanic/Latino, and 8% other. The experiment was approved by an IRB and all personally identifiable information was removed from the data.

2.2 Tasks and Measures

2.2.1 Conversation-Based Assessment Weather Task

The Weather CBA task [32, 34] is an innovative, computer-based task that engages students in simulated science inquiry around the topic of thunderstorms, including data collection, analysis and prediction, and justification of reasoning in the context of conversations with two virtual agents: Art, a virtual peer working on the task alongside the student, and Dr. Garcia, a scientist and authority figure guiding the students. Students place weather stations to collect simulated data on an impending thunderstorm, take notes from the data, and in a conversation-based item are presented with two of the notes (one of their notes, and one "created" by the virtual peer). Students are asked to explain which note should be kept for making predictions about the likelihood of a thunderstorm. Students should choose the note summarizing data across multiple weather stations (i.e., more data yields a better prediction).

Students were randomly assigned to one of two question framing conditions, which included different main questions posed to the student [34]. In both conditions, students were presented with the two notes and were asked to indicate which note should be kept for making predictions later. In the comparison-framing condition, the main question is posed as "Please look carefully at and compare these two notes. Which one do you think we should keep for making predictions later and why?" whereas in the alternate argumentation-framing condition, the question is posed as "Please look carefully at and compare these two notes. I think we should keep [your note/my note] for making predictions later. Do you agree with me? Why or why not?" In this argumentation-framing condition, the peer always pointed students to the "better" note summarizing data from multiple stations. Thus, the conversations included adaptivity based on the student's responses to the simulation components and their linguistic input.

2.2.2 Overall Task Performance

The Weather CBA task includes a combination of more traditional item types (multiple choice, constructed response) and technology-enhanced items (drag-and-drop, simulation items), in addition to simulated conversations characteristic of CBAs, with a maximum total score of 29 points.

2.2.3 Human Scores for Conversations

Responses were dichotomously scored by human raters along multiple dimensions (see Table 1), with a maximum of five possible points per conversation. Responses were scored both for the correctness of students' conclusions and the quality of the supporting reasoning, summed to create a single score for the entire response to the conversation (across all conversational turns). Raters were trained by a scoring leader to score these responses using a well-defined rubric. After two raters independently rated each category as 0 or 1, inter-rater reliability was examined. Initial inspection of the data revealed that there was an unequal cell distribution which can skew kappa results, so we only report percent agreement by dimension: Note Choice (93%); Immediate (90%); Relevant (82%); Sufficient (84%); and Aligned with Note Choice (88%).

Table 1. Dimensions of Human Scores and Definitions

Dimension	Definition
Note Choice	Students choose the note with more complete data represented (observations from multiple weather stations).
Reasoning: Immediate	Students provide reasoning immediately, within the first turn of the conversation.
Reasoning: Relevant	Students mention features related to the notes and the data they contain (e.g., weather stations, water vapor, instability).
Reasoning: Sufficient	Students mention that one note has more data than the other note.
Reasoning: Aligned with Note Choice	Students provide reasoning that is consistent with their choice of note.

2.2.4 LIWC

Linguistic Inquiry and Word Count (LIWC) is a computational linguistic tool that primarily uses a bag of words approach which identifies words in a given text and compares it to categories of words corresponding to parts of speech or at times broader constructs such as affect [24]. Overall, there are currently over 90 features that have shown to predict outcomes such as college GPA [25] to relationship longevity [16]. The system is available for research or for real time use, with an available API (https://www.receptiviti.com/liwc). In this study, we used a licensed desktop version of LIWC2015 [24] for post-hoc analysis of

the student contributions to the conversations with the artificial agents in the CBA task.

3. ANALYSES AND RESULTS

We began by analyzing all text with LIWC2015. Next, we inspected the data by question framing condition for sparseness (i.e., more than 95% of features with 0's), such that sparse features were removed. We also removed two outliers with an over-abundance of exclamation marks which were unique to these two students only (<1% of data lost). Next, we conducted a k-means cluster analysis on the remaining 34 features for 233 students. The results yielded four unique profiles discovered with a bottom-up approach. We interpreted the clusters via several methods including inspecting final cluster center means across features, qualitative analysis, and relationships to external variables including student demographics, question framing condition, human scoring of the conversations, and overall task performance.

3.1 Profiles

In total, 34 features were entered into a k-means clustering algorithm, which converged after 9 iterations. The results yielded 26 features with significant cluster centers (see Appendix A). The four clusters were titled Shallow Performers, One-Turn Wonders, Low Performers, and High Performers, with each described below.

3.1.1 Shallow Performers

The first profile was entitled "shallow performers" as these students were likely to give the correct answer to the question in terms of note choice (i.e., were above chance at picking the note that included more data) but did not provide good reasoning for their answers. The linguistic profile of these students as indicated in Appendix A included relatively higher levels of emotional tone (*M*=67, *SD*=38), positive emotions (*M*=8, *SD*=9), affect (*M*=9, *SD*=9) and the highest amount of dictionary words (*M*=90, *SD*=8), which is reverse coded indicating more words that are not in the dictionary, as well as a low word count (*M*=14, *SD*=8) compared to the other profiles. Qualitative analysis indicated that this group produced several responses in which students appeared to be implying that the peer agent is "smart" (e.g., argumentation-framing: "i do [agree] because you have good ideas._ I do, because you have good ideas and you sound smart."; comparison-framing: "We should keep Art's. Art's notes have more vocabulary and are more descriptive_ I don't know. Maybe because you are smarter."). This illustrates positive emotional tone with words such as "good" and "smart/smarter". The argumentation-framing responses earned credit for agreeing with peer's note choice but little credit for their explanations; comparison-framing responses were just above chance (60%) in note choice and also had quite poor reasoning.

3.1.2 One-Turn Wonders

Overall, these students provided "immediate" reasoning on the first turn of the conversation, but the likelihood of this reasoning being relevant was around chance levels; the reasoning was only rarely sufficient and aligned to note choice, although comparison-framing condition students scored slightly higher. Argumentation-framing condition responses were overwhelmingly likely to make the correct note choice, while comparison-framing responses were at chance levels. As seen in Appendix A, these students had the highest mean level of affect (*M*=10, *SD*=8), positive emotions (*M*=10, *SD*=8) and words per sentence (*M*=17, *SD*=11). From qualitative analysis, we saw that the first response included apparently longer sentences, consistent with providing immediate reasoning on turn one, and often used the word "better" (positive affect). An example argumentation-framing response is "Yes, you have a very good note. (I guess) _ It has a bunch of good information on it.". The student includes reasoning in the first turn (i.e., yes followed by attempted justification); with prompting, they provide additional detail ("good information") but no relevant reasoning. An example comparison-framing response is: "I personaly think my notes are better because i did notes for each seperate weather station. _ I had more hours to predict what happen and i wrote seperate notes for each weather station." This illustrates higher words per sentence and an attempt to provide immediate reasoning, but the note choice and reasoning are entirely incorrect. It appears these students understand the need to provide reasoning to support their note choice, but this reasoning was not relevant or sufficient.

3.1.3 Low Performers

These students performed poorly on all aspects of the task and human scores, with note choice at chance levels. Their linguistic profile as seen in Appendix A, showed a relatively low level of words per sentence (*M*=12, *SD*=8) which could indicate minimal attempts to provide a viable answer. These responses had high levels of personal pronouns (*M*=18, *SD*=10), perhaps an artifact of requiring students to say "my note" when choosing their own note. Qualitative analysis indicated that comparison-framing students often incorrectly chose "my note", and argumentation-framing students sometimes disagreed with Art. Their reasoning was more relevant than the shallow performers, but it was similarly insufficient and misaligned to note choice (although somewhat better for comparison-framing). This group also had more off-task responses, perhaps indicating disengagement. An example argumentation-framing response is "No because i did not give a water vaper percent _ no because i did not give a vaper number _ i did not give an exact answere", showing relevant but insufficient reasoning and incorrect note choice. A comparison-framing response with incorrect note choice and poor reasoning is "mine because im an actual person _ because im an actual person and not a computer program".

3.1.4 High Performers

These students performed relatively well on all human scores. As seen in Appendix A, these students had relatively high levels of words per sentence (*M*=15, *SD*=12) and analytical thinking (*M*=43, *SD*=30) which makes sense given the science inquiry task context. An example response is "yes because my note shows that there are no cold fronts _ There are no cold fronts" which includes the correct note choice, immediate reasoning on the first turn, and relevant (but not sufficient) reasoning (i.e., mentioning cold fronts) aligned with the note choice. Therefore, these students appear more likely to draw on relevant evidence within the task (above chance levels), with comparison-framing students especially likely to provide correct responses with more relevant reasoning.

3.2 Relationships to External Variables

Profiles were compared to external variables Specifically, we examine how features correlate with question-framing conditions and overall performance (see Appendix B), followed by analyses of the relationship between profiles to demographics. Relationships between profiles and question-framing conditions, human scores, and overall performance, were examined using a Kruskal-Wallis method with Monte Carlo simulation across 10,000 samples.

3.2.1. Profiles and Demographics

We conducted chi-square analyses between demographic variables and the four profiles. We discovered no significant differences leading us to interpret that profile membership was indeed diverse.

3.2.2 Profiles and Question-Framing Condition

There was a significant difference between profile membership and question-framing condition, (X^2 (3,233) = 8.07, p =.04, *partial* η^2 = .035). The Monte Carlo simulation for significance with 10,000 samples revealed a significance level of p = .04 (lower bound p =.038, upper bound p = .049). Mean ranks indicate that the Shallow Performers had the highest number of cases in the argumentation-framing condition (130.34) and the High Performers had the lowest (101.60). Although this may simply be an artifact of random assignment, it makes sense as students in the comparison-framing condition performed better overall. The other two profiles of One-Turn Wonders and Low Performers were in the middle with slightly higher mean ranks for the former over the latter (123.91 and 116.36, respectively).

3.2.3 Profiles and Human Scores

3.2.2.1. Note Selection

A significant relationship was discovered between the profiles and human scores for note selection (X^2 (3,233) = 10.106, p =.02, *partial* η^2 = .046). The Monte Carlo simulation for significance with 10,000 samples revealed a significance level of p = .02 (lower bound p =.014, upper bound p = .020). The mean ranks suggest the highest scores for the Shallow Performers (129.38) and the lowest for the Low Performers (98.51) with the One-Turn Wonders and High Performers having virtually equivalent mean ranks (121.59 and 121.98, respectively). These results are consistent with the fact that argumentation-framing students were overwhelmingly likely (~90%) to make correct note selections in both the Shallow and Low Performing profiles, vs. 50-70% for comparison-framing.

3.2.2.2. Immediate Reasoning

A non- significant relationship was not discovered between the four profiles and Immediate (X^2 (3,233) = 4.267, p = .234, *partial* η^2 = .018). Patterns revealed that One-Turn Wonders had the highest overall mean (126.64) though essentially similar to that for High Performers (122.21), but substantively different from both Shallow Performers (111.73) and Low Performers (108.51).

3.2.2.3. Relevant Reasoning

A significant relationship was discovered between the four profiles and relevant reasoning (X^2 (3,233) = 20.896, p <.001, *partial* η^2 = .089)). The Monte Carlo simulation for significance with 10,000 samples revealed a significance level of p <.001, (lower bound p =.000, upper bound p = .001). Mean ranks indicate that High Performers performed the best (142.19), whereas the Shallow Performers performed the worst (92.48). Not surprisingly, the One-Turn Wonders and Low Performers fell in the middle, with One-Turn Wonders showing higher scores (122.00 and 107.19, respectively). Relevant reasoning was most likely for High Performers and One-Turn Wonders in the comparison-framing condition.

3.2.2.4. Sufficient Reasoning

A significant relationship was discovered for the four profiles and sufficient reasoning (X^2 (3,233) = 13.974, p = .003, *partial* η^2 = .061). The Monte Carlo simulation for significance with 10,000 samples revealed a significance level of p =.003, (lower bound p =.001, upper bound p = .004). Mean ranks were highest for the High Performers (133.71) and lowest for the Shallow Performers (98.63) with One-Turn Wonders having higher scores than the Low Performers (124.38 and 109.05, respectively).

3.2.2.5. Supports Note Choice

A significant relationship was discovered for the four profiles and supporting note choice (X^2 (3,233) = 18.304, p = .001, *partial* η^2 = .084). The Monte Carlo simulation for significance with 10,000 samples revealed a significance level of p <.001, (lower bound p =.000, upper bound p = .001). Once again, mean ranks revealed High Performers had the highest score (139.99) and Shallow Performers had the lowest score (103.99). The One-Turn Wonders and Low Performers fell in the middle with higher scores for the One-Turn Wonders (115.21 and 105.44, respectively).

3.2.4 Profiles and Overall Task Performance

A significant relationship was discovered for the four profiles and overall CBA task performance (X^2 (3,233) = 11.332, p = .010, *partial* η^2 = .048). The Monte Carlo simulation for significance with 10,000 samples revealed a significance level of p <.001, (lower bound p =.007, upper bound p = .012). Mean ranks indicate the highest score for the High Performers (134.79) and the lowest for the Shallow Performers (98.41) with One-Turn Wonders still out performing Low Performers (126.94, 106.16, respectively).

3.3 Results and Conclusions

We discovered four profiles of students based on linguistic features identified with the computational linguistic tool LIWC. The resulting profiles consisted of Shallow Performers, One-Turn Wonders, Low Performers, and High Performers. The Shallow Performers had higher levels of emotional tone, especially positive affect words, but fewer dictionary words. The One-Turn Wonders were characterized by high positive affect words and high words per sentence, reflected in their longer attempted justifications for their choices. The Low Performers showed low words per sentence and high levels of personal pronouns, in part reflecting the design of the task, but also reflecting generally shorter responses with poorer quality of reasoning. Finally, the High Performers had relatively high levels of words per sentence and analytical thinking which has predicted GPA in previous studies [25]. These four profiles were then validated by external measures that revealed patterns in expected directions.

Given the IDEA conference theme and our diverse sample, we compared key demographic measures including school location, race/ethnicity, free/reduced price lunch status, and other factors to the four profiles and found no significant differences. This indicates that these linguistic profiles did not show different demographic composition even though these factors did relate to overall task performance in prior research [32]. The implications are that linguistic profiles may be a manner of detection and intervention that transcend demographics and therefore may enable greater inclusion during the learning experience. That said, we do acknowledge that our sample was predominantly White (about 80%) despite the large variation in other key variables. We also acknowledge the relatively small sample size and plan to attempt to replicate these findings on a larger data set from this same CBA task in future work.

In sum, these findings can guide the creation and augmentation of novel CBAs to support personalized learning based on students' interactions with the system, transcending demographic differences. The methodology employed may also inform other researchers' attempts to discover ways to adapt and personalize other learning and assessment environments based on students' use of language.

4. REFERENCES

[1] Atkinson, R. K. 2002. Optimizing learning from examples using animated pedagogical agents. *Journ. of Educ. Psychology*, 94,2, 412-427.

[2] Baylor, A. L., and Kim, Y. 2005. Simulating instructional roles through pedagogical agents. *Intern.l Journ. of Art.Intellige. in Educa.*, 15, 95-115.

[3] Bloom, B. S. 1956. Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive Domain. New York: McKay.

[4] Bennett, R. E., Persky, H., Weiss, A., and Jenkins, F. 2007. *Problem-Solving in technology rich environments: A report from the NAEP technology-based assessment project.* NCES 2007-466, U.S. Department of Education, National Center for Educational Statistics, U.S. Government Printing Office, Washington, DC.

[5] Biswas, G., Jeong, H., Kinnebrew, J., Sulcer, B., and Roscoe, R. 2010. Measuring self-regulate learning skills through social interactions in a teachable agent environment. *Research and Practice in Tech-Enhanced Learn, 5*, (Jul 2010), 123-152.

[6] Clarke-Midura, J., Code, J., Dede, C., Mayrath, M., and Zap, N. 2011. Thinking outside the bubble: Virtual performance assessments for measuring complex learning. In *Technology-based Assessments for 21st century skills: Theoretical and Practical Implications from Modern Research*, M. C. Mayrath, J. Clarke-Midura, & D. Robinson Eds., Information Age, Charlotte, NC, 125-147.

[7] Craik, F. I. M., and Lockhart, R. S. 1972. Levels of processing: A framework for memory research. *J. of Verb. Learn. and Verb. Beh.* 11,6, (Dec 1972) 671–684.

[8] Dillon, J. 1988. *Questioning and teaching: A manual of practice*. New York, NY: Teachers College Press.

[9] Forsyth, C. M., Luce, C., Zapata-Rivera, D., Jackson, G. T., Evanini, K., and So, Y. 2019. *Evaluating English language learners' conversations: Man vs. machine. Interna. J. on Comp. Assist. Lang. Learn. 32, 4 (May 2019), 398-417.* DOI= 10.1080/09588221.2018.1517126

[10] Gholson, B., Witherspoon, A., Morgan, B., Brittingham, J. K., Coles, R., Graesser, A. C., Sullins, J., and Craig, S. D. 2009. Exploring the deep-level reasoning questions effect during vicarious learning among eighth to eleventh graders in the domains of computer literacy and Newtonian physics. *Instruct. Science.* 37, 5 (Sep 2009), 487-493.

[11] Gil, L., Bråten, I. Vidal-Abarca, E., and Strømsø, H, 2010. Summary versus argument tasks when working with multiple documents: Which is better for whom? *Cont.y Educ. Psych.* 35, 3 (Jul 2010), 157-173. DOI= 35. 157-173. 10.1016/j.cedpsych.2009.11.002.

[12] Goldman, S. R., Duschl, R. A., Ellenbogen, K., Williams, S., and Tzou, C. T. 2003. Science inquiry in a digital age: Possibilities for making thinking visible. In *Cognition in a digital world*. H. van Oostendorp Ed. Erlbaum, Psychology Press. Mahwah, NJ, 253-284.

[13] Graesser, A. C. 2016. Conversations with AutoTutor help students learn. *Inter. J. of Artif. Intell. in Educ.* 26,1 (Mar 2016), 124-132.

[14] Graesser, A. C., Ozuru, Y., and Sullins, J. 2009. What is a good question? *In Threads of coherence in research on the development of reading ability*. M. G. McKeown, & L. Kucan, Eds. Guilford, New York, NY, 112-141.

[15] Graesser, A. C., and Person, N. K. 1994. Question asking during tutoring. *Amer. Educat. Res. J.* 31,1 (Mar 1994), 104–137.

[16] Ireland, M. E., Slatcher, R. B., Eastwick, P. W., Scissors, L. E., Finkel, E. J., and Pennebaker, J. W. 2011. Language style matching predicts relationship initiation and stability. *Psych. Sci.* 22,1, (Jan 2011), 39–44.

[17] Johnson, W. L., Rickel, J. W., and Lester, J. C. 2000. Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *Intern. J. of Artif. Intell. in Educ. 11,1 (Nov 2000), 47-78.*

[18] Johnson, W. L., and Lester, J. C. 2016. Face-to-face interaction with pedagogical agents, Twenty years later. *Inter. J. of Artif.l Intel. in Educ.* 26,1 (Mar 2016), 25–36.

[19] McNamara, D. S., O'Reilly, T., Best, R., and Ozuru, Y. 2006. Improving adolescent students' reading comprehension with iSTART. *J. of Educ. Comp. Res.*, 34,2 (Mar 2006), 147-171.

[20] Millis, K., Forsyth, C., Butler, H., Wallace, P., Graesser, A. C., and Halpern, D. 2011. Operation ARIES! A serious game for teaching scientific inquiry. *In Serious Games and Edutainment Applications*, M. Ma, A. Oikonomou, & J. Lakhmi Eds. Springer-Verlag, London, 169-196.

[21] Moreno, R., Mayer, R. E., Spires, H. A., and Lester, J. 2001. The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents? *Cog. and Instruct.*19, 2 (Jun 2001), 117-213.

[22] Mosenthal, P.1996. Understanding the strategies of document literacy and their conditions of use. *J. of Ed.Psych.* 88, 2 (June 1996), 314-332.

[23] Olney, A., D'Mello, S. K., Person, N., Cade, W., Hays, P., Williams, C., Lehman, B., and Graesser, A. C. 2012. Guru: A computer tutor that models expert human tutors. In *Proceedings of the 11th International Conference on Intelligent Tutoring Systems* (Springer, Berlin, Heidelberg, Jun 14, 2012), ITS2012. Springer-Verlag, Berlin,.256-261

[24] Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. 2015. *The Development and Psychometric Properties of LIWC2015*. University of Texas at Austin., Austin, TX. DOI: 10.15781/T29G6Z

[25] Pennebaker, J. W., Chung, C. K., Frazee, J., Lavergne, G. M., and Beaver, D. I. 2014. When small words foretell academic success: The case of college admissions essays. *PloS One*. 9, 12 (Dec 2014), 110 p. e115844.

[26] Quellmalz, E. S., Timms, M. J., Buckley, B. C., Davenport, J., Loveland, M., & Silberglitt, M. D. (2011). 21st Century Dynamic Assessment. In M.C. Mayrath, J. Clarke-Midura, & D. Robinson (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research*. Charlotte, NC: Information Age. 55-90.

[27] Reder, L. 1987. Strategy selection in question answering. *Cognitive Psychology*, 19, 1 (Jan 1987), 90- 138.

[28] Rouet, J. 2006. The skills of document use: From text comprehension to web-based learning. Mahwah, NJ: Erlbaum.

[29] Singer, M. 2003. Processes of question answering. In *Psycholinguistics*, G. Rickheit, T. Hermann, and W. Deutsch Eds. Walter de Gruyter, Berlin, 422-431.

[30] So, Y., Zapata-Rivera, D., Cho, Y., Luce, C., and Battistini, L. 2015. Using trialogues to measure English language skills. *J. of Educat. Tech. of Edu. Tech. and Society. Special Issue: Tech. Supp. Assess. in Formal and Inform, Learn*, 18, 2 (Apr 2015), 21-32. *Informal Learning* J. García Laborda, D. G Sampson, R. K. Hambleton and E. Guzman (Eds.). 21-32.

[31] Song, Y., Sparks, J.R., Brantley, W., Oliveri, M., and Zapata-Rivera, D. 2014. Designing Game Activities to Assess Students' Argumentation Skills. *Paper presented at the annual meeting of the American Educational Research Association (AERA)*, Philadelphia, PA.

[32] Sparks, J.R., Peters, S., Steinberg, J., James, K., Lehman, B.A., & Zapata-Rivera, D. (2019, April). Individual difference measures that predict performance on conversation-based assessments of science inquiry skills. In *Proceedings at the annual meeting of the American Educational Research Association* (Toronto, Canada, April 5-9, 2019).

[33] Sparks, J.R., and Zapata-Rivera, D. (2019, February). *Designing virtual agents for assessment.* Presentation to Educational Testing Service Constructed Response Forum, Princeton, NJ

[34] Sparks, J.R., Zapata-Rivera, D., Lehman, B., James, K., and Steinberg, J. 2018. Simulated Dialogues with Virtual Agents: Effects of Agent Features in Conversation-Based Assessments. In *Proceedings of International Conference on Artificial Intelligence in Education*, (Springer, Cham) 469-474. DOI= 10.1007/978-3-319-93846-2_88.

[35] Ward, W., Cole, R., Bolaños, D., Buchenroth-Martin, C., Svirsky, E., and Weston, T. 2013. My science tutor: A conversational multimedia virtual tutor. *J. of Educ. Psych.* 105, 4 (Nov 2013), 1115-1125.

[36] Zapata-Rivera, D., Jackson, T., Liu, L., Bertling, M., Vezzu, M., & Katz, I. R. (2014) Science Inquiry Skills using Trialogues. The *Proceedings of 12th International conference on Intelligence Tutoring Systems*. 625-626.

APPENDIX A

Table 2. Final Cluster Center Means and Standard Deviations

Feature	Clus1 (n=52)	Clus2 (n=50)	Clus3 (n=65)	Clus4 (n=66)	F
Word Count	14 (8)	23 (14)	19 (13)	21 (15)	4.6
Analytic Think	5 (6)	39 (28)	14 (22)	43 (30)	34.5
Emotional-Tone	67 (38)	95 (7)	35 (28)	24 (6)	104.2
Word per Sentence	11 (6)	17 (11)	12 (8)	15 (12)	4.3
Dictionary	90 (8)	86 (10)	88 (10)	77 (18)	14.2
function	58 (12)	49 (14)	60 (14)	47 (15)	14.1
personal-pronoun	20 (9)	10 (7)	18 (10)	7 (8)	30.9
“you”	10 (10)	3 (4)	3 (5)	1 (4)	23.0
“She/he”	1 (3)	1 (2)	0 (1)	2 (4)	4.4
article	2 (4)	5 (5)	3 (4)	5 (6)	7.0
prepositions	3 (5)	7 (6)	5 (5)	7 (7)	6.9
auxiliary verb	15 (8)	10 (6)	12 (9)	9 (6)	7.0
negations	1 (3)	1 (3)	11 (19)	2 (5)	12.9
verb	19 (10)	16 (7)	20 (12)	15 (9)	4.1
adjective	12 (9)	10 (9)	6 (7)	8 (8)	5.9
compare	9 (8)	8 (9)	4 (6)	7 (7)	5.3
number	0 (3)	2 (4)	1 (3)	2 (4)	2.9
affect	9 (9)	10 (8)	4 (7)	1 (3)	20.8
Positive emotion	8 (9)	10 (8)	2 (5)	0 (1)	32.6
male	1 (3)	1 (2)	0 (1)	2 (4)	4.2
insight	3 (6)	4 (6)	6 (6)	3 (4)	4.0
drives	10 (9)	9 (7)	4 (6)	2 (4)	19.5
achieve	3 (6)	4 (6)	0 (1)	0 (1)	9.8
reward	3 (6)	5 (7)	1 (2)	0 (1)	17.1
Focus past	1 (3)	1 (3)	3 (5)	1 (3)	4.8
Focus future	3 (5)	2 (3)	1 (2)	1 (3)	3.3

For a complete explanation of each variable, please see the LIWC manual 2015 [30]. The final resulting profiles mentioned above were titled Shallow Performers (Cluster 1), One-Turn Wonders (Cluster 2), Low Performers (Cluster 3) and High Performers (Cluster 4).

APPENDIX B

Table 3. Final Features Correlations to Total Score and Conditions

Feature	Total CBA Score	Question Framing
Word Count	.448**	-.117
Analytic Think	.159*	-.061
EmotionalTone	-.108	.184**
Word per Sentence	.327**	-.067
Dictionary	.033	.278**
function	.022	.160*
personal pronoun	-.122	.144*
“you”	-.225**	.146*
“She/he”	.034	-.251**
article	.148*	-.069
prepositions	.281**	.059
auxiliary verb	.155*	.011
negations	-.177**	.101
verb	.167*	-.010
adjective	.094	-.174**
compare	.097	-.278**
number	.085	-.187**
affect	-.275**	.249**
Positive emotion	-.263**	.217**
male	.040	-.253**
insight	.071	-.132*
drives	-.076	.142*
achieve	-.131*	-.086
reward	-.170**	.025
Focus past	.132*	.042
Focus future	.017	.177**

**p>.01, *p>.05