Interactive Listening—The Duolingo English Test

Duolingo Research Report DRR-23-01
April 25, 2023 (17 pages)
englishtest.duolingo.com/research

Geoffrey T. LaFlair*, Andrew Runge*, Yigal Attali*, Yena Park*, Jacqueline Church*, Sarah Goodwin*

Abstract

This paper introduces a new integrated task type on the Duolingo English Test called Interactive Listening and grounds the task within the Duolingo English Test's theoretical language assessment design framework and its assessment ecosystem. The task and automated item generation methods contribute to measurement of the constructs of L2 listening, reading, and writing, thereby strengthening the validity claims of the Duolingo English Test.

Keywords

Duolingo English Test, Interactive Listening, interactional competence, listening assessment

Contents

*Duolingo, Inc.

Corresponding author:
Geoffrey T. LaFlair
Duolingo, Inc. 5900 Penn Ave
Pittsburgh, PA 15206, USA
Email: englishtest-research@duolingo.com

1 Introduction

The Duolingo English Test is a digital-first, computer-adaptive, high-stakes proficiency test that assesses English language proficiency for admission to English-medium universities. The Duolingo English Test currently employs twelve different types of items to assess English proficiency in the academic context (Cardwell et al., 2022). Performance on these item types contributes to four subscores (Literacy, Conversation, Comprehension, and Production) and an Overall score. The test is designed to support both efficiency and effectiveness at all stages, from development, to administration, to scoring in large-scale standardized proficiency testing.

The new item type, Interactive Listening, complements the dictation item type on the DET. The dictation item type is an integrated test of listening and writing that requires test takers to utilize their aural processing skills and demonstrate comprehension by typing what they heard (Buck, 2001; Goodwin & Naismith, 2023). Interactive Listening builds on this by measuring aspects of interactional competence. It requires test takers to participate in a situationally driven conversation in a university setting. In this task, test takers listen to their interlocutors' turns, read and select the best response to each turn, and then summarize in writing the conversation they participated in. These tasks collectively enhance the construct coverage of the DET on dimensions of listening, reading, and writing in university settings. The description of the development of this task is framed in an e-ECD framework (Arieli-Attali et al., 2019; Langenfeld et al., 2022). Under this framework we discuss evidence for three models:

1. The e-Proficiency model, which describes the intended knowledge, skills, and abilities (KSAs) that the task measures;
2. The e-Task model, which describes the design of the task and process through which content, questions, keys, and distractors are generated;
3. The e-Evidence model, which describes the psychometric support for the item type.

These three models are supported by the DET's item development framework, which is illustrated in Figure 1. The Assessment planning / e-ECD stage consists of decisions about the domain, construct and task specifications for items. The ML/NLP modeling stage supports task generation, as it is in this stage that the features of the task (e.g., topics, and purposes for interacting) that can be used for scalable task creation are specified. In the human-in-the-loop review stage, subject matter experts review what is generated to ensure the tasks are construct-relevant (item quality review) as well as free of content that would be distracting to test takers (fairness and bias reduction review). Finally, the pre-piloting, large-scale piloting, and computational psychometrics stages are used to evaluate tasks and collect statistical evidence that supports the use of these tasks as a measure on the DET. Collectively, the processes illustrated in Figure 1 work together to support the three models in the e-ECD framework.

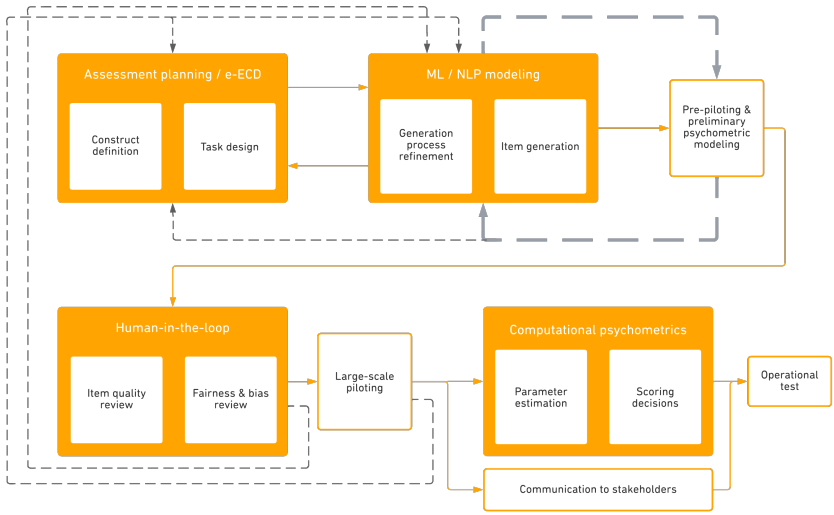

Figure 1. DET item development process

2 e-Proficiency model: Theoretical motivation for Interactive Listening

In this section, we describe the theoretical motivation for the interactive listening task, which is designed to measure characteristics of interactional competence. It allows stakeholders to make inferences about how well test takers can interact with peers and professors in university settings. It is an integrated listening, reading, and writing task, and consists of two stages. The first stage measures a test taker's ability to navigate conversations in academic settings with professors and with peers by participating in a simulated conversation. In the second stage, the test takers summarize, in writing, the conversation. This design creates a task that is a direct measure of test takers' receptive abilities, and an indirect measure of their spoken abilities in the construct of interactional competence, as well as a direct measure of their abilities to summarize in writing.

The construct of interactional competence has been conceptualized by Galaczi and Taylor (2018) as a tree. In this representation, the roots of the tree represent macro-level (speech situation) and micro-level features (speech event and speech act) that affect the characteristics of an interaction; the trunk represents the interlocutors and their relationship; and the branches of a tree represent interactional skills and subskills related to both speaking and listening in an interaction, such as topic management and turn-taking.

A tangible example of this in a university setting is an office hours conversation between a professor and a student. In this example, the speech situation is office hours, during which students discuss topics related to their academics with professors. The speech event is the actual conversation or discussion that the student has with their professor. Within this speech event are many speech acts (delivered by both interlocutors), including but not limited to greetings,

leave-takings, requests, and jokes. Examples of topic management include when (and if) the student nominates the topic of the discussion through a question or comment. Finally, an example of successful turn-taking would be if the student participates in the conversation by asking questions, sharing ideas, and even interrupting appropriately to make contributions to the conversation. When elements of the macro- and micro- level features vary (e.g, talking in a cafe with a peer about the same topic), the linguistic features of the skills categories can vary. It would be expected that the interaction with a peer would be more colloquial than the interaction with a professor (however, even this can vary!).

Variation in language use across contexts is an established claim in the applied linguistics research community, though approaches to theorizing and evaluating these differences can vary (c.f., Biber & Conrad, 2019; Canale & Swain, 1980; Dijk, 1977; Halliday & Matthiessen, 2004; Hymes, 1974). Incorporating this conceptualization of communication into assessment was made popular by Bachman (1990) and Bachman and Palmer (1996, 2010). For example, Bachman (1990) argued that a language test can be thought of as creating a specific context for someone to demonstrate their language skills. The elements of the test (like the types of tasks) are designed to simulate the kind of situations someone might encounter in the real world when using that language. Under this argument, the test designer manipulates characteristics of a target language use (TLU) domain (e.g., office hour conversation) so that they can design a task that is relevant to that TLU domain but that still fits within the context of a language test. Aspects of the TLU domain that are often considered for assessing interactional competence include who the participants are, their relationship to one another; the purpose, mode, and topic of communication; and characteristics of its setting.

One example of a task that measures interactional competence is the oral proficiency interview, or OPI (Galaczi & Taylor, 2018). The OPI is a face-to-face spoken interaction between an examiner and a test taker that seeks to establish the oral, or conversational, proficiency of a test taker. The purpose of the OPI is twofold: for the examiner to rate the test taker's proficiency and for the test taker to demonstrate proficiency. While the topics of conversation in an OPI will vary, they are typically determined by the examiner.

While OPIs are a useful indicator of speaking ability, they have been shown to be distinct from naturally occurring conversation and interviews in the TLU domain in terms of turn-taking, topic control, and question–response patterns, potentially threatening the ability to generalize performance on OPIs to performance in conversation in the TLU domain (Johnson & Tyler, 1998; Lier, 1989; Seedhouse, 1998; Young & He, 1998). These differences in linguistic characteristics between OPIs and conversations in the TLU domain stem from discrepancies between the situational characteristics of OPIs and the TLU domain, including differences in the purpose for communicating and in the power relation between interlocutors. As a result, while acknowledging the signal that OPIs give to test score users about a person's speaking ability, the extent to which this signal generalizes to conversation in any setting is relatively limited (Johnson & Tyler, 1998; Lier, 1989; Staples et al., 2017).

In addition to these aforementioned distinctions between conversations and OPIs, the real-time nature of the interview adds to the difficulty of evaluating test-taker performance. The administration of OPIs in the context of large-scale assessments is resource intensive with

respect to time, which increases the cost of administering these tasks. OPIs require a substantial amount of training and quality control activities to ensure inter-rater reliability as well as OPI administration comparability (Chalhoub-Deville & Fulcher, 2003). Because of these difficulties, tasks measuring interactional competence, including OPIs, are rarely included in large-scale standardized assessments of English proficiency. IELTS is an example of a large-scale standardized assessment with an OPI portion (Weir et al., 2013). Other large-scale standardized assessments of English proficiency either do not include tasks measuring interactional competence (e.g., TOEFL iBT) or include limited or indirect measures. For example, the TOEFL Essentials requires test takers to listen to a conversation and either answer comprehension questions about the conversation or complete the last turn of the conversation by selecting from several options (Papageorgiou et al., 2021).

3 e-Task model: The Interactive Listening task

In this section we present the design of the task, as well as the processes for generating the content, questions, keys, and distractors for the task. The Interactive Listening task that we present in this paper is designed to address some of the limitations of existing assessment tasks. The Interactive Listening task requires test takers to engage in a conversation with an interlocutor (see Appendix for details). Test takers are provided with a scenario to set the stage for the conversation; they listen to the interlocutor's turns, which are presented by an animated character that stands in for the interlocutor; they then respond by selecting the best response from several text options; they are provided with immediate feedback about which option was the best response, which is important for helping test takers stay on track in the conversation. This cycle is repeated 5-6 times through the conversation's conclusion. Finally, after the conversation is over, test takers are asked to summarize in writing the conversation they participated in.

The Interactive Listening task demonstrates correspondence to the TLU domain via interlocutors and reasons for communication. The interlocutors in the Interactive Listening task are peers and professors. The reasons for communicating with those interlocutors are asking for clarification about lecture content, making requests, gathering information, asking for advice, planning study sessions, and participating in other university-oriented conversations (Biber & Conrad, 2019). The topics, purposes, and turns are fixed in these conversations, rendering the task limited in its ability to fully measure turn management or topic nomination.

In particular, the task includes three types of conversations: student-student conversations that focus on requests, advice seeking, and other university-oriented purposes; student-professor conversations that focus on similar purposes; and student-professor conversations that focus on information seeking where the student needs to get information about a specific topic from their professor. The professor's lines in these conversations are significantly longer, with the goal of requiring stronger listening skills from the test taker to process the information.

The tasks' interlocutors and their voices are drawn from the cast of the Duolingo World Characters (Chiu et al., 2021; Hartman, 2020). Their names are Bea and Oscar and they both play the role of peer and professor across the different conversation types. All of the listening input is delivered via a text-to-speech (TTS) system developed for the Duolingo World Characters.

The Interactive Listening task is the first task in a large-scale standardized assessment to measure how well a test taker can navigate a full conversation. The situational context of the task aligns with real-world situations for interacting with peers and professors in university contexts with respect to the topics, the reasons for the conversation, and the relationship between interlocutors. This is further supported by the authentic nature of the turns that test takers listen to and select when completing the task. Finally, the summary task adds additional measurement of how well the test takers understand the conversation and how well they can summarize in English, an important skill for writing in university settings.

3.1 Automatic generation approach

We use the GPT-3 (Brown et al., 2020) family of models to generate conversations which can be used for both items (individual turns) and distractors. GPT-3 takes as input a text prompt, which is formatted as a set of instructions for what types of content the model should generate as well as a small number of examples demonstrating the task, which can provide additional information about the structure and style of the desired output. This method of prompting a large language model for output following a specific format has been explored for generating various types of content such as dialogues and narratives in recent years (Lee et al., 2022; Yang et al., 2022), and it was used by Attali et al. (2022) for generating passages, questions, and distractors for an interactive reading task.

3.1.1 Conversation generation

Conversations are generated in a three step process that starts with the curation of topics and conversational goals by subject matter experts on the test development team. We start with a list of 130 basic scenarios describing conversational goals or topics between two students or between a student and a professor. These scenarios cover a wide range of interactive tasks such as requesting help, asking for advice, providing feedback or recommendations, seeking information, discussing and comparing options, and describing a recent experience. We then use GPT-3 to add details to these scenarios, such as specific classes, assignments, or academic subjects. These additional details allow us to create richer and more engaging conversations that are aligned with the TLU domain.

Finally, we use GPT-3 to generate conversations based on these detailed scenarios. We use examples of conversations paired with a scenario description to prompt GPT-3 to generate new conversations based on new scenarios. The examples used depend on the conversation type we wish to generate: conversations between two students, conversations between a student and a professor, or detailed conversations between a student and a professor on a specific topic.

Conversations were selected in such a way that they were evenly distributed by their base scenario and the academic subjects they covered. We selected 900 of the generated conversations to turn into items based on the following criteria:

- 5-6 turns per speaker
- Test taker will have the final line in the conversation
- Test taker's lines are approximately the same length within and across item sets
- No potentially inappropriate or offensive words or phrases

3.1.2 Multiple choice item generation For each conversation, one of the student roles (or the only student role) is assigned to the test taker. We create a multiple choice question for each of that speaker's turns where the real line of dialogue is the correct answer.

Extending the work of Attali et al. (2022), we create incorrect answers (distractors) to multiple choice questions (turns) by selecting turns from other conversations in the pool of generated conversations, amounting to over one million turns of dialogue. Appropriate distractors from other conversations are selected based on a wide range of criteria:

- Similarity with the base scenario of the conversation
- Similarity with the detailed scenario of the conversation
- Whether or not the line is spoken by a speaker in a student role
- Whether the line occurs at the same approximate point in the conversation - near the beginning, middle, or end
- Ratio of the length of the distractor to the correct answer
- Similarity to the correct answer
- Average similarity to other candidate distractors
- Similarity to distractors used in other questions
- Difference in the probability of generating this distractor as the next line in the conversation compared to the correct answer

3.1.3 Conversation summaries Finally, we generate summaries of the conversations to use as references when grading the conversation summarization task. We use GPT-3 to summarize the conversation both from a first-person perspective as a participant in the conversation and also from a third-person perspective to ensure test takers can respond from either perspective without potentially impacting their scores.

4 e-Evidence model: Psychometric support

This section presents statistical evidence that supports the task, which was collected during the pilot phases of task development. The interactive listening task was developed through an iterative process that included multiple pilots that examined the effects of task design decisions on psychometric performance of the tasks. As such, this process exemplifies the computational psychometrics approach (von Davier et al., 2021), which blends data-driven computer science methods (machine learning and data mining, in particular) and theory-driven psychometrics and language assessment design framework considerations in order to measure latent abilities in real time. This blend is often instantiated as an iterative and adaptive process (Burstein et al., 2022; Langenfeld et al., 2022; von Davier, 2017).

4.1 Initial pilots

The piloting platform used for this process was developed as part of the DET practice test. This practice test is a short version of the DET, freely available online to anyone interested in the test, with over ten thousand test sessions daily. The piloting platform is an opt-in section at the end of the practice test with experimental tasks; around 50% of practice test takers choose to complete these additional experimental tasks. We used this platform to conduct around 10

controlled experiments over the course of six months to test different task features as well as conduct surveys of current and former test takers to collect their feedback on the task. In this section, we briefly report the results of several of these experiments and surveys to demonstrate the process.

These experiments investigated the effects of the following on test taker performance and psychometric properties of items:

- The number of multiple choice options
- The availability of the conversation history
- The number of allowed plays of the audio
- Varying lengths of allotted time
- Various design and UI decisions (such as the position of the play button)

In addition to these larger scale experiments, we also conducted a small-scale pilot with 18 current undergraduate and graduate students that included in-depth questionnaires about participants' experiences with the task. The students took two Interactive Listening tasks, with follow-up survey items asking whether they liked the Interactive Listening tasks and felt the tasks assessed their listening, interactional, and writing skills. Survey items also inquired what the participants thought of the animated characters, what test takers' perceptions of receiving implicit feedback were, and whether they encountered any technical difficulties. In general, participants' feedback was overwhelmingly positive. We used their feedback to iterate on the task design.

The preliminary data that was collected in these data collections contributed to the decisions made regarding task features and provided support for the Interactive Listening task. Scores on the Interactive Listening task showed moderate correlations with other item types on the test; they also showed moderate correlations with self-reported scores on other large-scale high-stakes standardized English proficiency tests.

4.2 Large scale pilot

The culmination of the iterative task development process was a large-scale pilot, whose purpose was to evaluate the quality of the AIG processes described above, both from a human review and psychometric perspective.

For this pilot, around 1,000 conversations were generated. All items underwent item quality, fairness and bias (FAB), and audio quality reviews. These reviews were conducted by twenty-five external reviewers and six internal Duolingo team members. Reviewers had diverse backgrounds with regard to gender identity, age, and racial/ethnic background. All had at least a bachelor's degree (and in some cases a Ph.D.) in linguistics, language studies, or a related field. All had expertise in teaching and assessing in a relevant linguistic and cultural context.

Each scenario, conversation, and item first went through a two-phase item quality review process. In phase one, reviewers ensured that the scenarios provided sufficient context for the test taker to successfully engage in the conversation. For the conversations, reviewers evaluated the cohesion, clarity, and logical consistency throughout the text, confirming that each conversation

included a speech situation, a speech event, and a speech act. For questions, reviewers judged the viability of each option by ensuring that the correct answer was correct and the distractors were incorrect. Item reviewers, who were trained on fairness guidelines, edited out any content that seemed to violate these guidelines. All proposed edits were reviewed by a second set of reviewers, who carefully evaluated, revised and incorporated the proposed changes into the final items. If an item set required extensive edits that would take beyond twenty minutes, then reviewers abandoned the item set.

Following assessment fairness guidelines, FAB reviewers reviewed the remaining scenarios, conversations, and questions for any content that was either potentially controversial, too culturally specific, or potentially unfamiliar to the intended global test taker audience. Then all the TTS-generated conversation turns underwent an audio quality review. Turns were revised where there was inappropriate pronunciation, word and sentence-level stress, and intonation.

In summary, following all reviews and adjudication, approximately 80% of original conversations were retained for the large-scale pilot. Overall, each conversation was reviewed by 6-7 people and the review process took about one hour per conversation across all reviews.

The large-scale pilot was conducted on the practice test. On average, 464 responses were collected per item. The items showed a wide distribution in difficulty and in most cases medium to high item-total correlations. Further analyses were performed to remove distractors with lower discrimination indices to improve the overall discriminatory power of the items.

The results of the pilots of Interactive Listening have demonstrated that these items have met the minimum requirement for subsequent, more complex psychometric modeling for inclusion in the Conversation and Comprehension subscores by demonstrating (1) their association with other measures and (2) their psychometric qualities.

5 Test-taker readiness materials and practice tests

The DET delivers updates about any changes to the test that would impact the test taker experience well ahead in advance. In addition, an extensive readiness guide and unlimited practice tests are available at no cost to ensure that test takers' performance are free of bias due to unfamiliarity of test tasks and response format. The DET also communicates updates on the test via various media channels including YouTube, Facebook, and Twitter.

6 Discussion

The evidence-centered design of the Interactive Listening task helps build evidence for the digitally-informed chain of inferences for using DET scores for their intended purposes (Arieli-Attali et al., 2019; Burstein et al., 2022; Langenfeld et al., 2022), particularly pertaining to domain descriptions and scoring. We describe how the design of the Interactive Listening task creates an interaction between the test taker and an interlocutor that aligns with aspects of the construct of interactional competence. The digital-first nature of the DET, including the mode of delivery for the input and the response format, helps fulfill the digital consideration of authentic and construct-relevant interaction on the test. The automated generation process allows for the

creation of large sets of tasks that represent the variety of situations, topics, and purposes for communicating with peers and professors in the university. The evidence specification activity shows that responses from Interactive Listening are indicative of test taker listening, reading, and writing skills.

7 References

Arieli-Attali, M., Ward, S., Thomas, J., Deonovic, B., & von Davier, A. A. (2019). The expanded evidence-centered design (e-ECD) for learning and assessment systems: A framework for incorporating learning goals and processes within assessment design. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00853

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & von Davier, A. A. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, *5*. https://doi.org/10.3389/frai.2022.903077

Bachman, L. F. (1990). *Fundamental considerations in language assessment*. Cambridge: Cambridge University Press.

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice*. Oxford University Press.

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice*. Oxford University Press.

Biber, D., & Conrad, S. (2019). *Register, genre, and style* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781108686136

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 1877–1901). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Buck, G. (2001). *Assessing listening*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732959

Burstein, J., LaFlair, G. T., Kunnan, A. J., & von Davier, A. A. (2022). *A theoretical assessment ecosystem for a digital-first assessment: The Duolingo English Test* (Duolingo Research Report DR-22-03 DRR-22-01). Duolingo.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1–47.

Cardwell, R., LaFlair, G. T., & Settles, B. (2022). *Duolingo English Test: Technical manual*. Duolingo. https://duolingo-papers.s3.amazonaws.com/other/det-technical-manual-current.pdf

Chalhoub-Deville, M., & Fulcher, G. (2003). The oral proficiency interview: A research agenda. *Foreign Language Annals*, *36*(4), 498–506.

Chiu, E., Lenzo, K., & Swecker, G. (2021). *Giving our characters voices*. https://blog.duolingo.com/character-voices/

Dijk, T. A. van. (1977). *Text and context: Explorations in the semantics and pragmatics of discourse*. Longman.

Galaczi, E., & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly*, *15*(3), 219–236. https://doi.org/10.1080/15434303.2018.1453816

Goodwin, S., & Naismith, B. (2023). *Assessing listening on the Duolingo English Test* (Duolingo Research Report DR-23-03 DRR-23-03). Duolingo.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2004). *An introduction to functional grammar* (3rd ed.). Routledge. https://doi.org/10.4324/9780203783771

Hartman, G. (2020). *Building character: How a cast of characters can help you learn a language*. https://blog.duolingo.com/building-character/

Hymes, D. (1974). *Foundations in sociolinguistics*. University of Pennsylvania Press.

Johnson, M., & Tyler, A. (1998). Re-analyzing the OPI: How much does it look like natural conversation? In R. Young & A. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 27–52). John Benjamins.

Langenfeld, T., Burstein, J., & von Davier, A. A. (2022). Digital-first learning and assessment systems for the 21st century. *Frontiers in Education*, *7*. https://doi.org/10.3389/feduc.2022.857604

Lee, Y.-J., Lim, C.-G., Choi, Y., Lm, J.-H., & Choi, H.-J. (2022). PERSONACHATGEN: Generating personalized dialogues using GPT-3. *Proceedings of the 1st Workshop on Customized Chat Grounding Persona and Knowledge*, 29–48.

Lier, L. van. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly*, *23*(3), 489–508. http://www.jstor.org/stable/3586922

Papageorgiou, S., Davis, L., Norris, J. M., Garcia Gomez, P., Manna, V. F., & Monfils, L. (2021). *Design framework for the TOEFL essentials test 2021* (RM-21-03). ETS. https://www.ets.org/Media/Research/pdf/RM-21-03.pdf

Seedhouse, P. (1998). Oral proficiency interviews as varieties of interaction. In S. Ross & G. Kaspar (Eds.), *Assessing second language pragmatics* (pp. 199–219). Palgrave Macmillan.

Staples, S., LaFlair, G. T., & Egbert, J. (2017). Comparing language use in oral proficiency interviews to target domains: Conversational, academic, and professional discourse. *The Modern Language Journal*, *101*(1), 194–213. https://doi.org/https://doi.org/10.1111/modl.12385

von Davier, A. A. (2017). *Computational psychometrics in support of collaborative educational assessments*.

von Davier, A. A., Mislevy, R. J., & Hao, J. (2021). Introduction to Computational Psychometrics: Towards a Principled Integration of Data Science and Machine Learning

Techniques into Psychometrics. In *Computational Psychometrics: New Methodologies for a New Generation of Digital Learning and Assessment* (pp. 1–6). Springer.

Weir, C. J., Vidakovic, I., & Galaczi, E. D. (2013). *Measured constructs: A history of cambridge english language examinations 1913-2012* (Vol. 37). UCLES/Cambridge University Press.

Yang, K., Tian, Y., Peng, N., & Klein, D. (2022). Re3: Generating longer stories with recursive reprompting and revision. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, 4393–4479.

Young, R., & He, A. (1998). *Talking and testing: Discourse approaches to the assessment of oral proficiency*. John Benjamins.

A Appendix

Live demonstrations of a student-student conversation and a student-professor conversation are available. The task starts with a scenario that describes who the test taker is talking with and for what purpose.

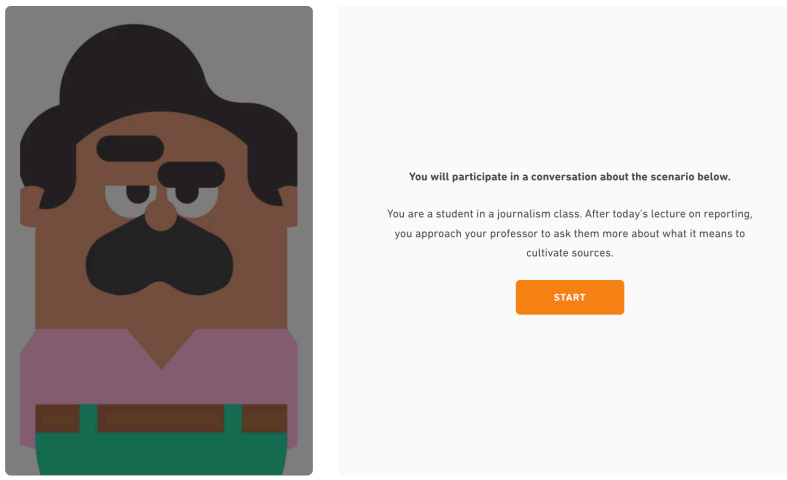

Figure 2. Task scenario

When the test takers click on the start button, the conversation begins. In this task, the test taker makes the opening turn of the conversation. In other tasks, the interlocutor will have the first turn.

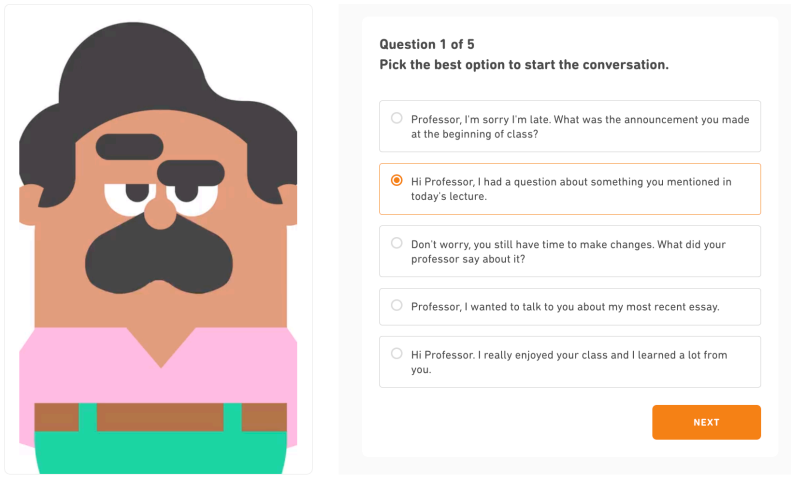

Figure 3. Opening turn

When the test taker selects the best answer for the turn, they receive visual feedback through color (green) and a check mark. On the first turn for the interlocutor, test takers are reminded that they can only listen to the turns in the conversation once.

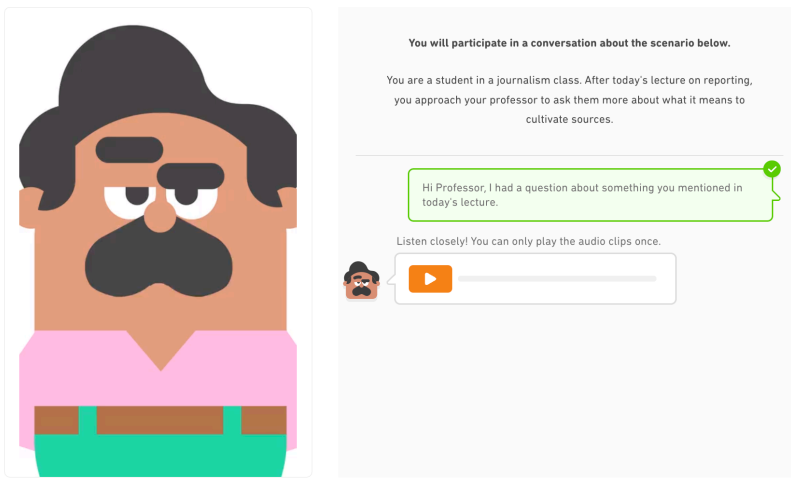

Figure 4. Correct-answer feedback and play button

Test takers receive visual feedback when they select an incorrect option. The color of the box turns light red; there is an "x" in the upper right corner of the dialogue box; their answer is struck through, and the best answer is presented.

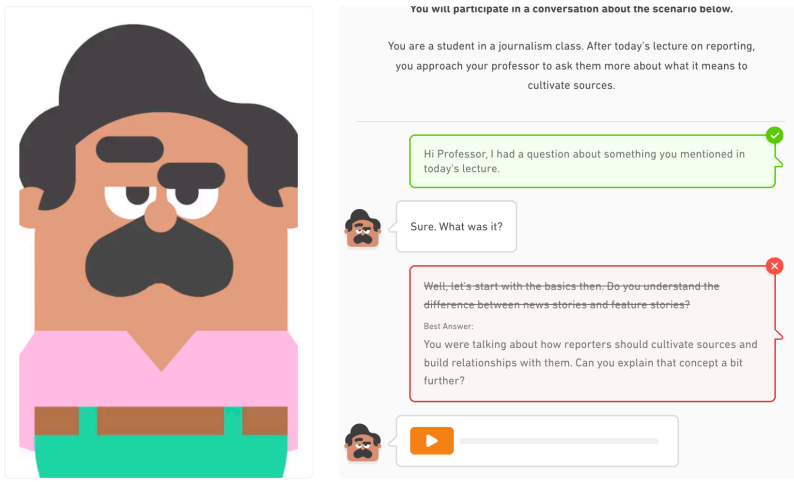

Figure 5. Incorrect-answer feedback

This process repeats itself until the conversation is over. Test takers are notified that the task is complete. If they have time left, they can review the conversation before clicking next to move on to the summary task.

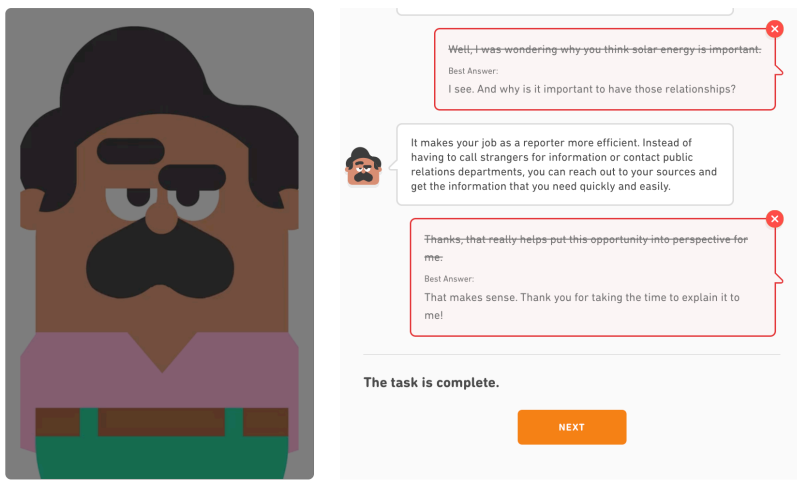

Figure 6. Task completion