Measuring the Power of Learning.®

Research Memorandum
ETS RM–19-01

Designing an End-of-Unit Assessment for a Computer-Based Pragmatics Learning Tool: First Steps Toward a Prototype Module

Heidi Liu Banerjee

Veronika Timpe-Laughlin

February 2019

ETS Research Memorandum Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Memorandum series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Memorandum series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Designing an End-of-Unit Assessment for a Computer-Based Pragmatics Learning Tool: First Steps Toward a Prototype Module

Heidi Liu Banerjee
Teachers College, Columbia University, New York, New York

Veronika Timpe-Laughlin
Educational Testing Service, Princeton, New Jersey

February 2019

Corresponding author: V. Timpe-Laughlin, E-mail: VLaughlin@ets.org

Suggested citation: Banerjee, H. L., & Timpe-Laughlin, V. (2019). *Designing an end-of-unit assessment for a computer-based pragmatics learning tool: First steps toward a prototype module* (Research Memorandum No. RM-19-01). Princeton, NJ: Educational Testing Service.

Action Editor: Donald Powers

Reviewers: Ikkyu Choi and Tanner Jackson

Abstract

This paper reports on the design and development of an end-of-unit assessment module for a self-access learning platform that is intended to help second and foreign language learners (L2) of English become (more) aware of pragmatic phenomena in the U.S. workplace. After providing a brief rationale, we first describe the computer-based, interactive pragmatics learning tool called *Words at Work*. Following the evidence-centered design (ECD) framework, we then outline the design of the prototype end-of-unit assessment module for Unit 3, a unit that deals with the speech act of request in general workplace communication. After presenting the proficiency model, the evidence model, and the task model, we provide in detail an overview of the task and item designs, including feedback implementation and ideas for scoring. We conclude the report with a discussion of the insights gained during the development process, as well as an outlook on future directions for research and development efforts.

Key words: *Words at Work*, workplace pragmatics, request, assessing pragmatics, evidence-centered design

Table of Contents

List of Tables

Page

List of Figures

Pragmatic knowledge has been widely recognized as an integral component of communicative language competence because it involves a person's ability to construe and convey nuanced meanings that vary contextually, sociolinguistically, socioculturally, psychologically, and rhetorically (Grabowski, 2009, 2013; Purpura, 2004, 2016). As noted by Crystal (1997), at the center of pragmatics is how participants perceive and use language, including "the choices they make, the constraints they encounter in using language in social interaction and the effects their use of language has on other participants in the act of communication" (p. 301). In other words, utterances in successful communication are both context-dependent and interlocutor-sensitive.

Because pragmatics is shaped by a variety of linguistic and non-linguistic factors, researchers over the years have attempted to exhaustively define what constitutes pragmatics (e.g., Leech, 1983; Levinson, 1983; Yule, 1996) and how pragmatics situates in a communicative language ability model (e.g., Bachman, 1990; Bachman & Palmer, 1996; Purpura, 2004). Synthesizing how pragmatic competence has been conceptualized and operationalized over time, Timpe-Laughlin, Wain, and Schmidgall (2015) proposed that pragmatic competence draws upon sociocultural knowledge, pragmatic-functional knowledge, grammatical knowledge, discourse knowledge, and strategic knowledge, and that meaning co-construction heavily relies on interactions between these different types of knowledge within and between the interlocutors.

The complexity of pragmatics has led to many challenges in learning, teaching, and assessment in the second or foreign language (L2) context. From the learning perspective, in order to demonstrate pragmatic competence, L2 learners need to have relevant L2 knowledge, skills, and abilities, and also understand and have the ability to apply "non-linguistic rules of social conduct and social relationships" when appropriate (Roever, 2009, p. 561). From the teaching perspective, material selection and teachers' pedagogical knowledge as well as pedagogical content knowledge can all pose potential instructional challenges (Cohen, 2008). And finally, from the assessment perspective, while the approaches to assessing pragmatics have evolved in the past few years, from paper-and-pencil-based discourse completion tasks to context-rich, technology-enhanced interactive tasks, the challenge remains that there has been a lack of discursive-orientedness in the approaches (Grabowski, 2016). Therefore, while the teaching, learning, and assessing of L2 pragmatics have received much attention in the past few years (Purpura, 2016), there is still a lot to be done to ensure that the instruction, materials, and assessment properly reflect the learning needs.

One of the target language use (TLU) domains where pragmatics plays a crucial role is workplace settings, particularly because globalization of the economy has brought together working professionals of diverse language and cultural backgrounds, and English has mostly been adopted as the medium of communication. However, for working professionals who come from a non-English-speaking background, communicating their messages accurately and appropriately can sometimes be challenging. Whether it is due to their lack of linguistic resources to express themselves appropriately or their unfamiliarity with the target social or cultural norms, pragmatic failures—when compared to grammatical errors—are more likely to create undesirable impressions about the speaker (Taguchi & Sykes, 2013; Washburn, 2001). In fact, pragmatic failure has been identified as a major cause of communication breakdown in workplace environments (Clyne, 1994). Thus, there is a need to provide non-English-speaking working professionals with sufficient learning opportunities to minimize undesirable high-stakes consequences.

In order to address the learning, teaching, and assessment needs and challenges in promoting L2 pragmatic competence, specifically in the TLU domain of workplace settings, Educational Testing Service (ETS) is developing a computer-based, self-access pragmatics learning tool, *Words at Work*. Focusing on the assessment component of the learning tool, this report delineates the design principles as well as the task and item specifications of an end-of-unit assessment of one learning module of *Words at Work.* The main purpose is for this end-of-unit assessment to serve as an assessment prototype and for this report to serve as a documentation to guide future development of assessment components within *Words at Work.*

***Words at Work*—The Pragmatics Strategy Builder**

The computer-based learning platform *Words at Work*,was developed to address the resource gap in the teaching and learning of L2 pragmatics. It is designed as a self-access, interactive learning platform in which pragmatics in workplace interactions is explicitly instructed through interactive tasks and simulated scenarios for adult English language learners who have at least an intermediate level of proficiency in English (approximately CEFR B1+) (Timpe-Laughlin et al., 2015). The goal of *Words at Work* is to develop adult L2 learners' pragmatic knowledge to facilitate their communicative language proficiency and to ultimately raise their awareness of pragmatic phenomena in interactions, specifically in workplace settings. Guided by the design principles and considerations identified in Timpe-Laughlin (2016), *Words at Work* is designed to simulate the interrelated steps of a real-life job cycle in a U.S.-English-speaking workplace, starting with a job hunt, followed by a job interview, the first day on the new job, up through the development of a regular job routine. Embedded in this scenario structure are nine learning modules, each of which focuses on a specific pragmatic phenomenon or speech act that is important for successful communication in the workplace, such as making requests, suggestions, and small talk. The scenario frame provides a narrative structure that is intended to contextualize the interaction for the learner while providing as much authenticity as possible. Figure 1 illustrates the scenario structure of *Words at Work.*

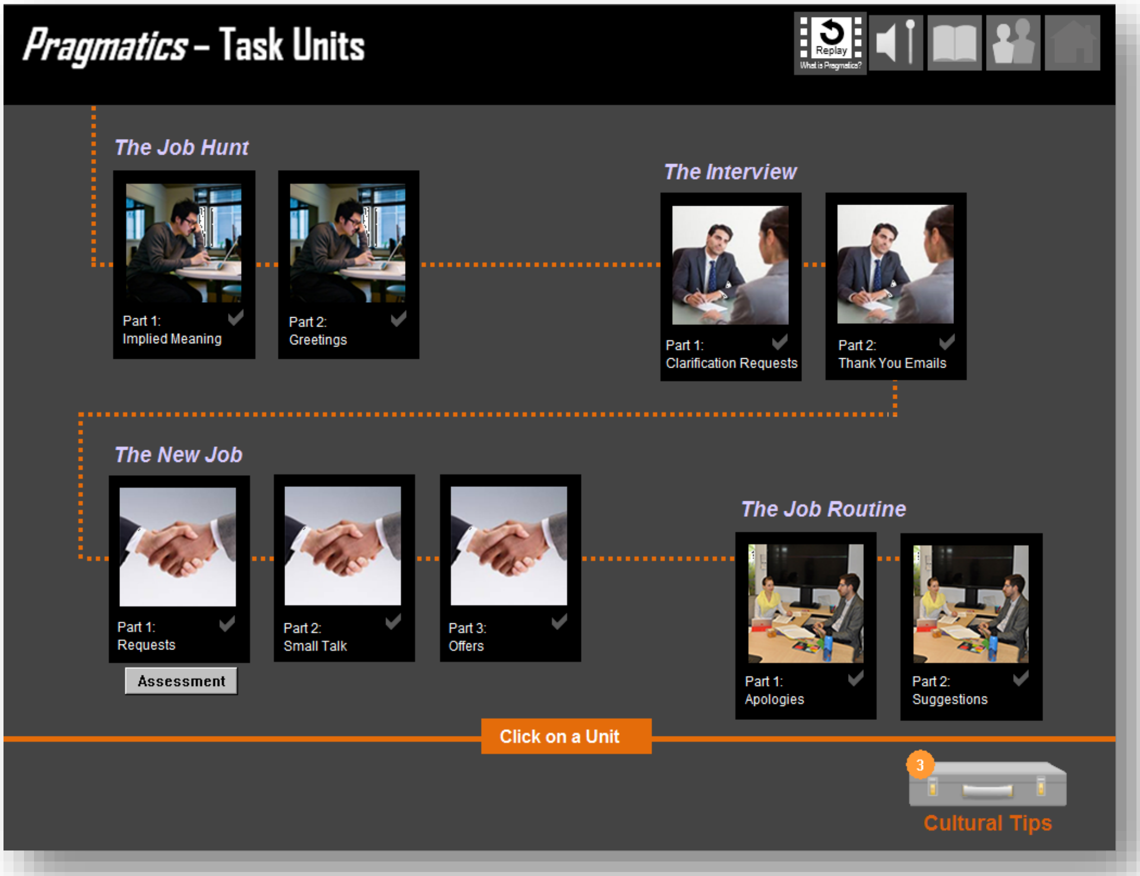


Figure 1. Main menu, featuring scenario structure of *Words at Work*.

The pragmatic content and tasks in each unit are informed by the pragmatic construct proposition and domain analysis of language use in workplace contexts of Timpe-Laughlin et al. (2015). Accordingly, each learning module is divided into three conceptual parts (i.e., a warm up and introduction, a sociopragmatic section, and a pragmalinguistic section), and is structured to offer the learner explicit input in the form of a series of e-lectures targeting sociopragmatic and pragmalinguistic features of a particular speech act, interspersed with interactive tasks, while progressing from receptive to productive skills.

In the warm up and introduction section, the learner is presented with a warm-up task to introduce the topic while setting up the narrative structure in order to more effectively immerse the learner within the scenario. Then a brief lecture is given to introduce the speech act being discussed and its relevance to the workplace, followed by a set of interactive tasks. Next, in the sociopragmatic section, the learner is presented with sociopragmatic-based lectures that focus on relevant sociocultural and contextual factors such as age, social distance, and degree of power relations between the interlocutors. These lectures are followed by tasks designed to enhance the observational and receptive skills of the learner on these key contextual variables. The learning modules then shift to a more pragmalinguistic focus with lectures identifying the semantic formulae and pragmalinguistic strategies necessary to appropriately use the given speech act, followed by further interactive tasks that become progressively complex and productive in nature. Roughly modeled after Bloom's (1956) taxonomy, the units are structured to progress from receptive to productive skills to promote observation, reflection, and production in the learner and to more directly target noticing as well as increased sensitivity and awareness on the part of the learner. As learners progress through the learning modules, they interact with a number of different characters, including colleagues, superiors, and friends. These recurring characters make it possible to control for and focus on key contextual features, such as who the interlocutor is, and thus, to parse the units and modules in a way that rewards learner progress (i.e., a character can go from a stranger to an acquaintance to a friend) while still maintaining coherence within a unit. For example, depending on the context of communication, it may be more appropriate to speak and/or respond to a particular character less formally across units, even though their relative status may not change.

Development of the End-of-Unit Assessment

As a self-access learning tool, it is important for *Words at Work* to provide its users with ample opportunities to monitor and evaluate their learning progress; therefore, an end-of-unit assessment was envisioned to be embedded in each learning module in the fully functional version of *Words at Work*. To ensure that the assessment not only allows learners to monitor their learning progress but also facilitates learning when appropriate, the prototype assessment module in *Words at Work* prioritizes learners' learning goals, formative evaluation, and feedback in the assessment process.

To reiterate, the purpose of this report is to illustrate how a computer-based, learning-oriented assessment module for a self-access pragmatics learning tool was conceptualized and designed. As a first prototype, this assessment module focuses on the speech act of request which is covered primarily in Unit 3[1] in *Words at Work.* The rationale behind selecting this particular unit to develop the prototype assessment module was because at the time of assessment development, this module contained the most complete learning materials; therefore, it was deemed the most appropriate unit to develop an end-of-unit assessment. In the following sections, the construct definition, the test design framework, the task types, and the item specifications of the assessment module of the request unit are described.

Construct Definition: Targeted Ability

Because the assessment module in *Words at Work* is designed to be a unit test that evaluates the extent to which learners have mastered the learning objectives, it is important for the assessment to reflect what learners are expected to learn from the unit. Therefore, the measured construct in each unit is defined as the targeted pragmatic ability of the unit. Based on the learning objectives of Unit 3 in *Words at Work,* learners' pragmatic knowledge of requests is measured by their ability to identify the nature of a request (i.e., direct, indirect, or hint), identify the purpose of a request (e.g., to ask for information, to set up a meeting, to ask for assistance), demonstrate their understanding of the social context when making a request (i.e., formal or informal), and understand and use proper linguistic strategies when making different types requests.

Test Design Framework

With the construct definition established, the next step is to define a test design framework based on which tasks and items can be systematically developed. For the purpose of having a coherent structure of test design, the evidence-centered design (ECD) framework (Mislevy, Steinberg, & Almond, 1999) was adopted. Based on the ECD framework, a structured test development process systematically starts with domain analysis, followed by domain modeling, conceptual assessment framework, assessment implementation, and concludes with assessment delivery. Through this process, ECD "ensures that the way in which evidence is gathered and interpreted is consistent with the underlying knowledge and purposes the assessment is intended to address" (Mislevy, Almond, & Lukas, 2003, p. 2).

Within the ECD framework, the most relevant component to this assessment development report is the conceptual assessment framework (CAF), which describes the operational elements of an assessment. It can, as Mislevy and Riconscente (2005) put it, be viewed as "machinery for generating assessment blueprints by means of a structure that coordinates the substantive, statistical, and operational aspects of an assessment" (p. 16). Thus, the CAF framework provides specifications to answer the fundamental questions regarding the operationalization of an assessment, including "What are we measuring?" (the proficiency model), "How do we measure it?" (the evidence model), "Where do we measure it?" (the task model), and "How much do we need to measure?" (the assembly model) (Mislevy, Almond, & Lukas, 2003). In the present study, we focused on the first three questions and their corresponding models.

The proficiency model. The proficiency model (or the student model) describes the variables "related to the knowledge, skills, and abilities we wish to measure" (Mislevy et al., 2003, p. 6). In this study, the proficiency model corresponds to the defined construct of the assessment module of *request*.

The evidence model. The evidence model describes "the observable behaviors or observable products of behavior resulting from responses to a particular task" (Zieky, 2014, p. 83). In order for learners' observable behaviors (i.e., assessment performance) to yield meaningful and interpretable validity claims, it is essential to explicate how these observable variables are connected to the proficiency model.

Based on the learning objectives in Unit 3 in *Words at Work*, we identified the following list of evidence to be observed in learners' performance in the self-assessment module:

1. Learners can correctly identify requests.
2. Learners can correctly identify the purpose of requests.
3. Learners can correctly determine the appropriateness of a request.
4. Learners can make requests to fulfill specific purposes (e.g., arrange a meeting, ask for a report).
5. Learners can make requests in written and spoken communication with proper structure and language.
6. Learners can employ various linguistic strategies when making requests given the social distance (i.e., social relationships between the interlocutors based on familiarity), relative power (i.e., power relationships between the request-initiator and the request-recipient), and degree of imposition (i.e., how big or difficult the request is).

The evidence model allows us to ensure that the assessment tasks represent the learning objectives. By mapping each task with one or more of the piece of evidence in the list, learners' performance outcomes in the end-of-unit assessment can be properly interpreted in terms of their mastery of the learning materials in the unit.

The task model. After the evidence to be observed is established, the task model is used to "describe how to structure the kinds of situations we need to obtain the kinds of evidence needed for the evidence models" (Mislevy et al., 2003, p. 10). In the assessment module of *requests*, four task types were identified to fulfill the purpose of this self-assessment. Table 1 briefly summarizes each task.

Table 1. Summary of the Tasks in the Assessment Module of Requests

Task	Item type	Focused skills	Evidence observed: *Learners will be able to…*
Task One	SR	Receptive skills	• identify requests (#1) • identify the purpose of requests (#2) • determine the appropriateness of a request (#3)
Task Two	Likert-scale type	Receptive skills	• identify the purpose of requests (#2) • determine the appropriateness of a request (#3) • make requests to fulfill specific purposes (#4)
Task Three	SR + CR	Receptive & production skills (limited; writing & speaking)	• determine the appropriateness of a request (#3) • make requests to fulfill specific purposes (#4) • make requests in written and spoken communication with proper structure and language (#5) • employ various linguistics strategies when making requests given the social distance, relative power, and imposition (#6)
Task Four	CR	Receptive & production skills (extended; speaking)	• make requests to fulfill specific purposes (#4) • make requests in written and spoken communication with proper structure and language (#5) • employ various linguistics strategies when making requests given the social distance, relative power, and imposition (#6)

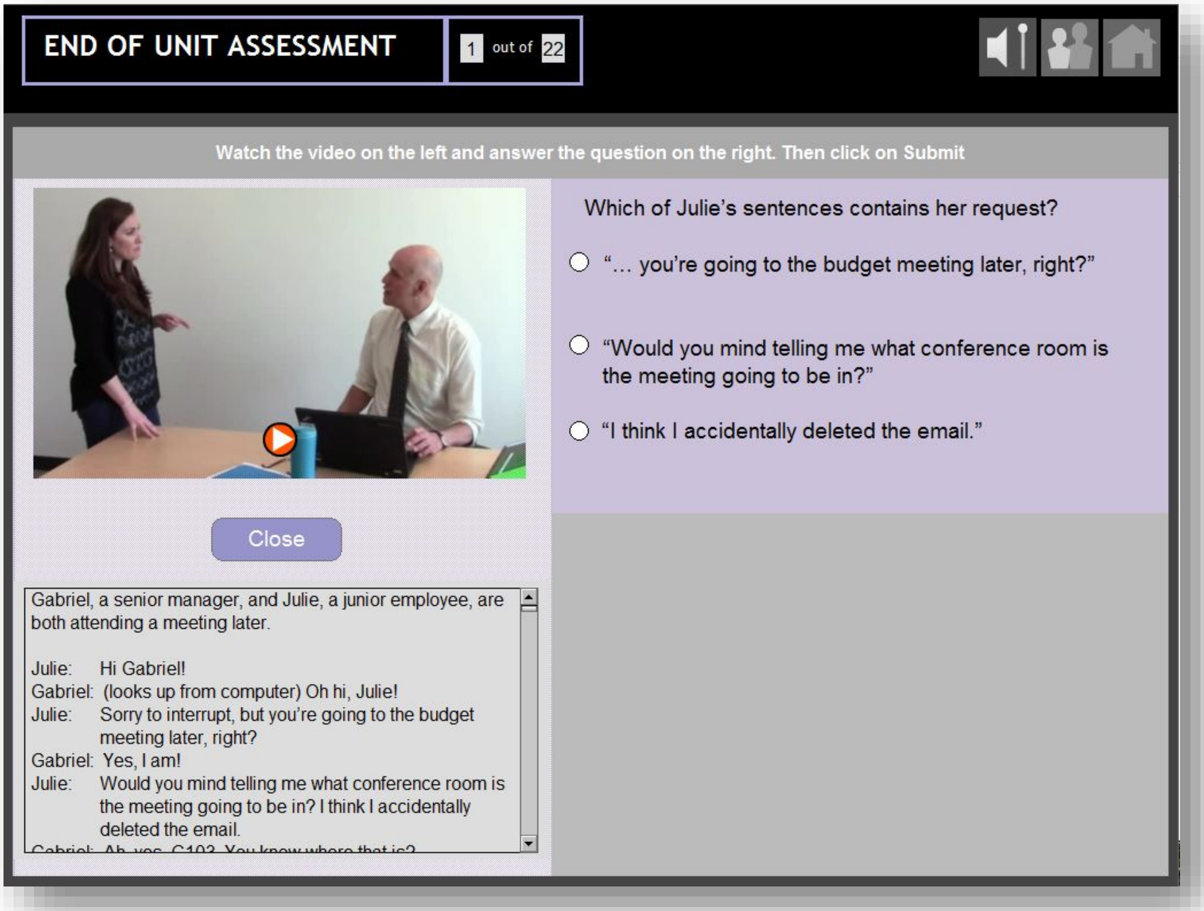
Note. SR = selected response; CR = constructed response.

In alignment with the learning module, the sequence of the assessment tasks also roughly follows Bloom's (1956) taxonomy. In other words, the tasks are structured in a way that learners can first demonstrate their understanding of the pragmatics of *requests* with familiar tasks (i.e., tasks that are similar to the learning tasks in the units) before moving to less familiar tasks to demonstrate their ability to transfer and apply their pragmatic knowledge.

Task Overview

With the ECD-informed test design framework of the assessment module established, the following section provides an overview of each task. Overall, we designed four tasks that aim to collect evidence of learners' knowledge, abilities, and skills of identifying and making requests. Each task has a number of items, with each item connected to a different scenario. Because the design of the scenarios is specific to the items, each item is embedded with a set of questions so that different aspects of pragmatic knowledge related to the same scenario can be measured.

Task One was designed to assess whether learners can identify the utterance that constitutes a request in a dialogue, its purpose, and its appropriateness in relation to the context. In this task, each item set contains the instruction, a description of the context and the characters, a short video, and four questions. The instruction, which states "Watch the video. When you are ready, go to the next page to answer some questions. You may watch the video twice," is the same across all items in Task One. The rationale for allowing learners to watch the video prompts twice is twofold. First, interactions in real time are not rewindable; therefore, allowing learners to watch the video without any watch-time restrictions may not allow us to interpret their pragmatic knowledge in interactions accurately. On the other hand, given that learners have to apply their pragmatic knowledge in a scenario where limited contextual clues are given, and it is not our intention to have learners memorize the conversation in order to perform the task, we decided to allow them to watch the video twice. Further, given the learning-oriented nature of this assessment module, a transcript of the conversation featured in the video is built into the task for scaffolding purposes. That is to say, learners who feel that they need additional assistance may click on the display button to deliberately access the assistance. Figure 2 illustrates a sample item in Task One.



Watch the video on the left and answer the question on the right. Then click on Submit

Which of Julie's sentences contains her request?

○ "… you're going to the budget meeting later, right?"

○ "Would you mind telling me what conference room is the meeting going to be in?"

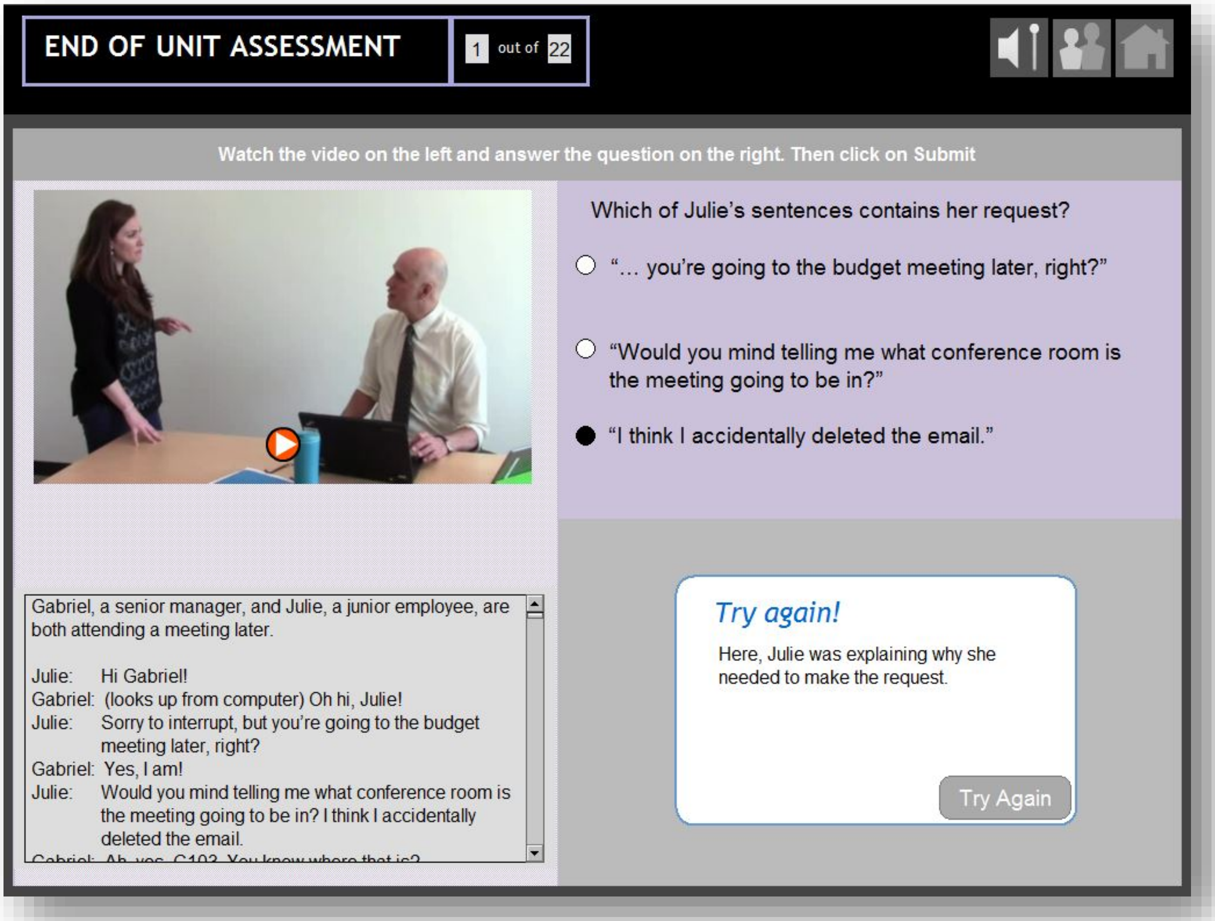
○ "I think I accidentally deleted the email."

Gabriel, a senior manager, and Julie, a junior employee, are both attending a meeting later.

Julie: Hi Gabriel!
Gabriel: (looks up from computer) Oh hi, Julie!
Julie: Sorry to interrupt, but you're going to the budget meeting later, right?
Gabriel: Yes, I am!
Julie: Would you mind telling me what conference room is the meeting going to be in? I think I accidentally deleted the email.
Gabriel: Ah, yes, C103. You know where that is?

Figure 2. Sample item from the request identification task (Task One).

After learners watch the video, they are presented with a series of four questions that aims to measure different aspects of pragmatic knowledge related to identifying requests. For each question, learners are provided with different types of feedback depending on their answer, and the feedback ranges from confirmation of a correct answer ("That's correct!"), to scaffolded assistance ("Try again!…"), to explicit correction ("Actually, the correct answer is…"). Figure 3 shows an example of the scaffolded "Try again!" prompt when learners answer a question incorrectly.



Watch the video on the left and answer the question on the right. Then click on Submit

Which of Julie's sentences contains her request?

○ "… you're going to the budget meeting later, right?"

○ "Would you mind telling me what conference room is the meeting going to be in?"

● "I think I accidentally deleted the email."

Gabriel, a senior manager, and Julie, a junior employee, are both attending a meeting later.

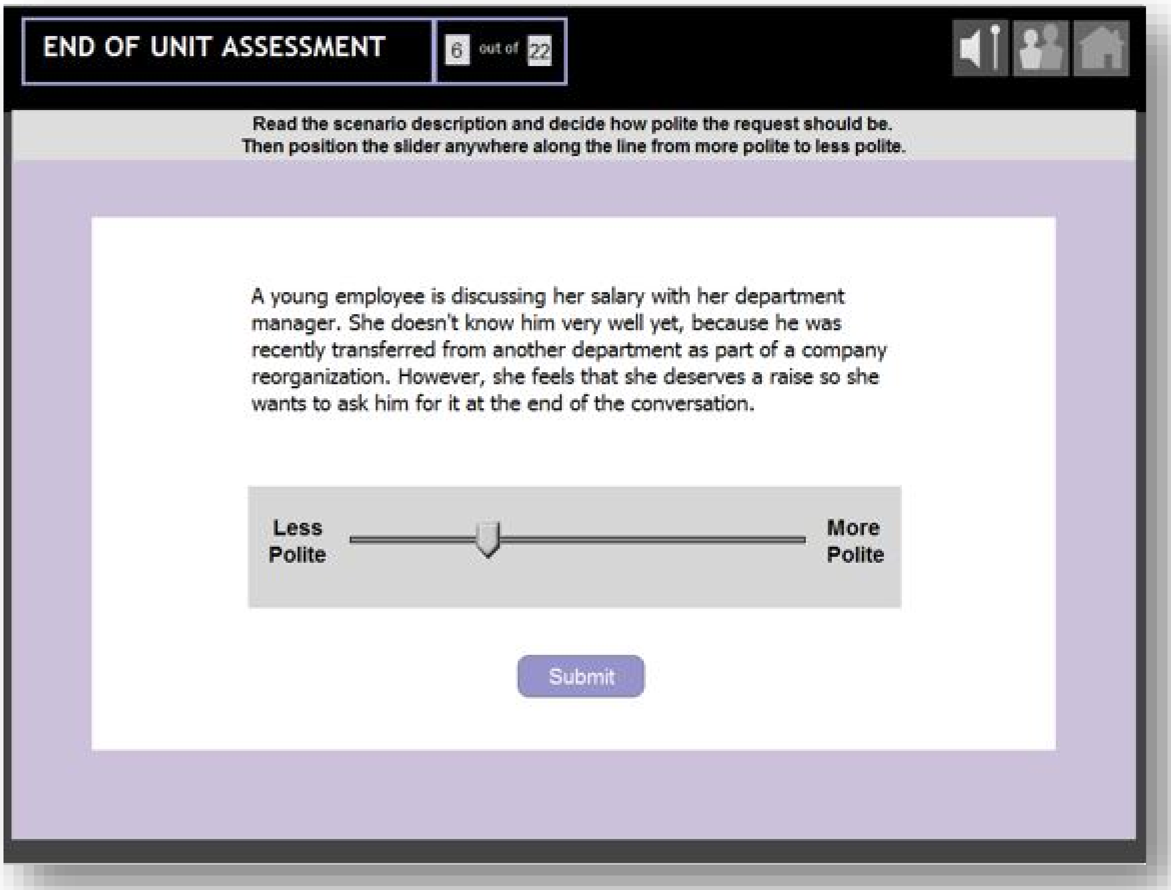
Julie: Hi Gabriel!
Gabriel: (looks up from computer) Oh hi, Julie!
Julie: Sorry to interrupt, but you're going to the budget meeting later, right?
Gabriel: Yes, I am!
Julie: Would you mind telling me what conference room is the meeting going to be in? I think I accidentally deleted the email.
Gabriel: Ah, yes, C103. You know where that is?

Try again!

Here, Julie was explaining why she needed to make the request.

Figure 3. Sample "Try again!" prompt and assistance for incorrect responses in the request identification task (Task One).

Task Two, the politeness judgment task, is a Likert-scale type task that aims to elicit learners' judgment on how polite a speaker should be when making requests based on the contextual variables. Learners are instructed to read the description of a scenario, and then indicate how polite the wording of the request should be on a slider from *less polite* on one end to *more polite* on the other. Figure 4 shows a sample item in Task Two. Conceptual ideas regarding the evaluation of responses to these items (i.e., feedback and scoring processes) are provided in the Item Specifications section below.



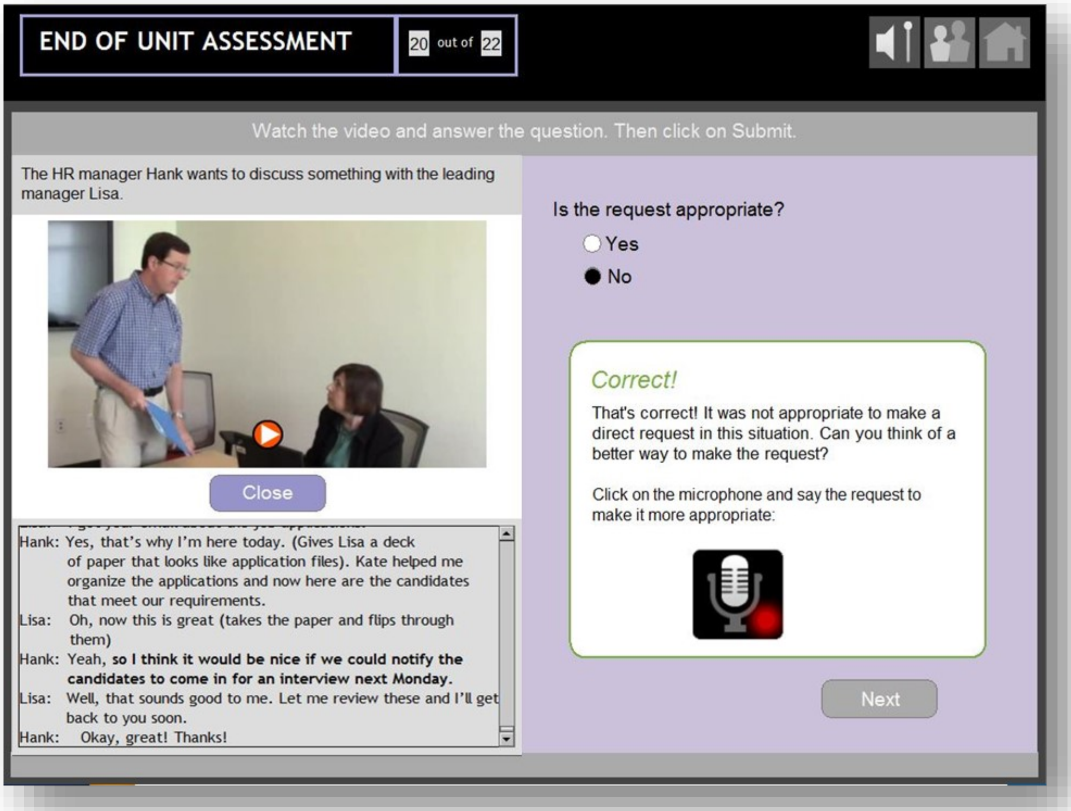
Read the scenario description and decide how polite the request should be.
Then position the slider anywhere along the line from more polite to less polite.

A young employee is discussing her salary with her department manager. She doesn't know him very well yet, because he was recently transferred from another department as part of a company reorganization. However, she feels that she deserves a raise so she wants to ask him for it at the end of the conversation.

Figure 4. Sample item from the politeness judgment task (Task Two).

Task Three, an appropriateness judgment task, aims to elicit both receptive and productive skills. Learners are asked to either read an e-mail (Part One) or watch a video (Part Two) and decide whether the request embedded in the e-mail or the dialogue is appropriate in the given scenario. If the request is correctly identified as inappropriate, learners are then prompted to write down (for e-mail items), or record (for video items) the appropriate alternatives.

For each item, learners are presented with brief but concrete instructions (e.g., "Read the e-mails and answer the questions"), a description of the context and the characters, an e-mail or a video, and the pragmatic judgment questions. Depending on learners' answers, different types of feedback are designed. If the request is appropriate, and learners also indicate that it is appropriate, they are given a positive feedback "That's correct!" and a brief explanation. If learners indicate that it is not appropriate, they are given a detailed explanation of why the request is, in fact, appropriate. If the request is not appropriate, and learners identify it correctly as inappropriate, they are prompted to write down or record what they would say to be appropriate. After they submit their answer, a list of model responses will be provided for learners to compare their answer with. However, if learners fail to identify the request as inappropriate, they will see a detailed explanation of why the response is, in fact, inappropriate. Then, they are asked to think of and provide an alternative response to make the request appropriate. After they submit their answer, they will see the model responses. The purpose of such a task design is not only for us to observe whether learners can apply what they have learned by using their pragmatic knowledge to analyze the appropriateness of the requests, but also to provide learners with feedback to enhance learning and promote awareness. Figure 5 shows a sample speaking item and its feedback design.



Watch the video and answer the question. Then click on Submit.

The HR manager Hank wants to discuss something with the leading manager Lisa.

Hank: Yes, that's why I'm here today. (Gives Lisa a deck of paper that looks like application files). Kate helped me organize the applications and now here are the candidates that meet our requirements.

Lisa: Oh, now this is great (takes the paper and flips through them)

Hank: Yeah, **so I think it would be nice if we could notify the candidates to come in for an interview next Monday.**

Lisa: Well, that sounds good to me. Let me review these and I'll get back to you soon.

Hank: Okay, great! Thanks!

Is the request appropriate?

○ Yes

● No

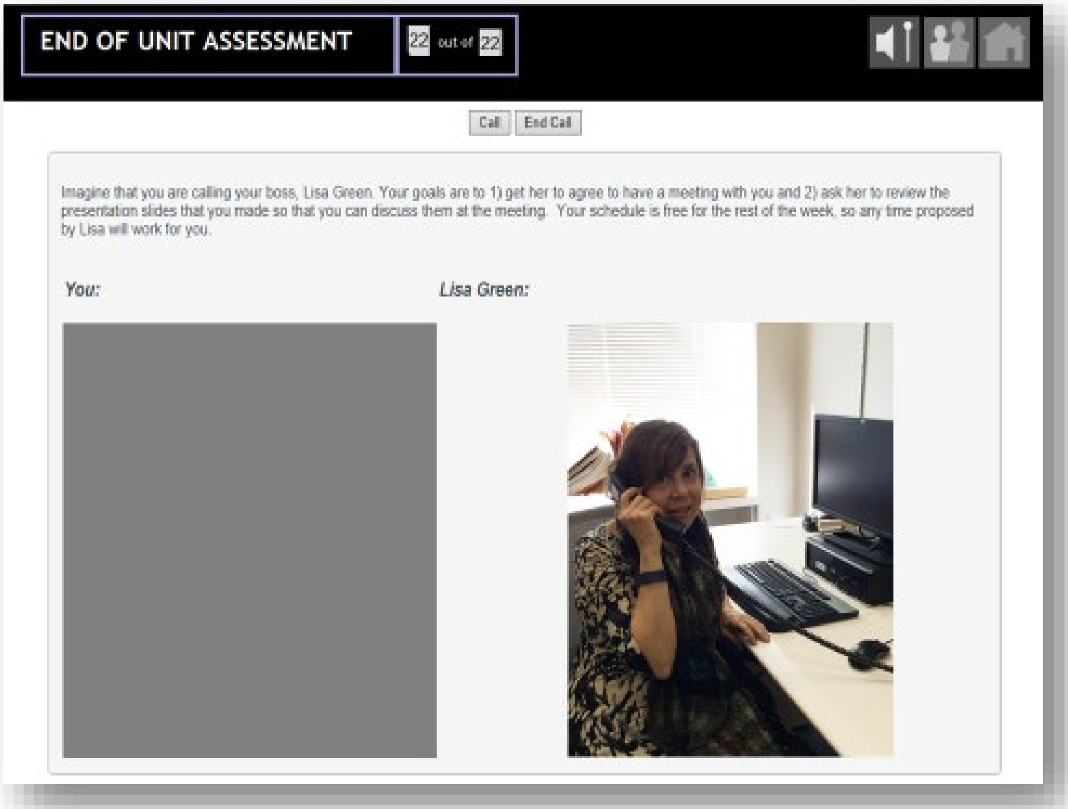
Correct!

That's correct! It was not appropriate to make a direct request in this situation. Can you think of a better way to make the request?

Click on the microphone and say the request to make it more appropriate:

Figure 5. Sample item from the appropriateness judgment task (Task Three).

Finally, Task Four, a multi-turn speaking task, was designed as a conversation-based speaking task, utilizing the HALEF spoken dialogue system and natural language processing (NLP) techniques (Ramanarayanan et al., 2017). Through automatic speech recognition and semantic understanding, learners can interact with a simulated agent through phone calls via the Web. Because it allows learners to produce their spoken language extensively, it provides further evidence regarding whether learners are able to apply and transfer their knowledge to a simulated, "real life" situation. Figure 6 shows a sample item layout of Task Four. Similar to commercial video-based telephony systems (e.g., Skype), learners see themselves in the grey box shown under the heading You in Figure 6. The learner's interaction with the item is thus audio- and video-recorded, using their computer's in-built audio-visual capabilities. For a more detailed understanding of the development of this item type, see Timpe-Laughlin et al. (2017).



Imagine that you are calling your boss, Lisa Green. Your goals are to 1) get her to agree to have a meeting with you and 2) ask her to review the presentation slides that you made so that you can discuss them at the meeting. Your schedule is free for the rest of the week, so any time proposed by Lisa will work for you.

You:

Lisa Green:

Figure 6. **Sample item layout from the multi-turn speaking task (Task Four).**

Item Specifications

This section specifies the details of how items in each task were developed. The procedures were intended to serve as the recommended item-writing protocol for item-bank expansion.

Task One

Task One, the request identification task, is a selected response task that includes a context statement, a conversation between two interlocutors, four multiple-choice question sets and their corresponding feedback. The purpose of Task One is to measure learners' ability to identify how requests are made in various workplace scenarios.

Context statement. The context statement briefly describes the setting to introduce the background of the interactions in the video. For example, the context statement for Item 1 in Task One, "Gabriel and Julie are both attending a meeting later," allows learners to understand that the conversation between Gabriel and Julie is about the meeting. If the characters belong to the set of key characters in the learning module (e.g., Lisa Green, Alex Robinson), there is no further description of their relationship in the context statement because it is assumed that learners are familiar with these recurring characters. However, if a new character is introduced, his/her relationship with the interlocutor is described.

Conversation. The conversation is between two interlocutors, one of whom makes a request (i.e., request-initiator) to the other (i.e., request-recipient). Given that we intend to measure whether learners can identify the request embedded in the conversation, it is essential to have a multi-turn conversation, with each turn serving a different pragmatic function. In order to provide sufficient contextual clues while at the same time minimizing cognitive load, the length of the conversation is approximately six turns (i.e., three pair parts) in our design. The formality of the situation, social distance, relative power, and degree of imposition (i.e., size of the request) are not ambiguous.

Question sets. Each item in Task One has four questions. Question 1 asks which of the request initiator's sentences contains his/her request, Question 2 asks what the purpose of the request is, Question 3 asks whether the situation is formal or informal, and Question 4 asks what type of request—direct, indirect, or hint—the request-initiator has used. These questions are designed to correspond to the structure and the content of the units in *Words at Work*. For example, the questions for Item 1 include:

1. Which of John's[2] sentences contains his request?
2. What was the purpose of the request?
3. Is the situation formal or informal?
4. What kind of request did John use?

Each question (except for Question 3) has three options: a key and two distractors. The options for Questions 1 and 2 vary, depending on the scenario; the options for Questions 3 and 4, however, are fixed.

Feedback. For each option learners select, a predesigned feedback appears to inform learners whether the answer is correct or incorrect, and why or why not that is the case. For the correct answers, the feedback prompt begins with "That's correct!" followed by a brief explanation for knowledge reinforcement. For the incorrect answers on first try, the feedback prompt begins with "Try again" followed by a hint for assistance. It is important that the hint doesn't give the answer away. The hint also varies depending on which wrong answer learners choose. If learners still answer incorrectly after the hint, they are provided with the correct answer and a more detailed explanation (i.e., "Actually, the correct answer is… because..."). Note that for Question 3, there are only two options (formal or informal), so learners will not be asked to try again if they answer incorrectly on first try. Such a feedback loop design, as illustrated in Figure 7, allows us to use explicit feedback to enhance learning and thus ensure the learning-oriented feature of the assessment module.

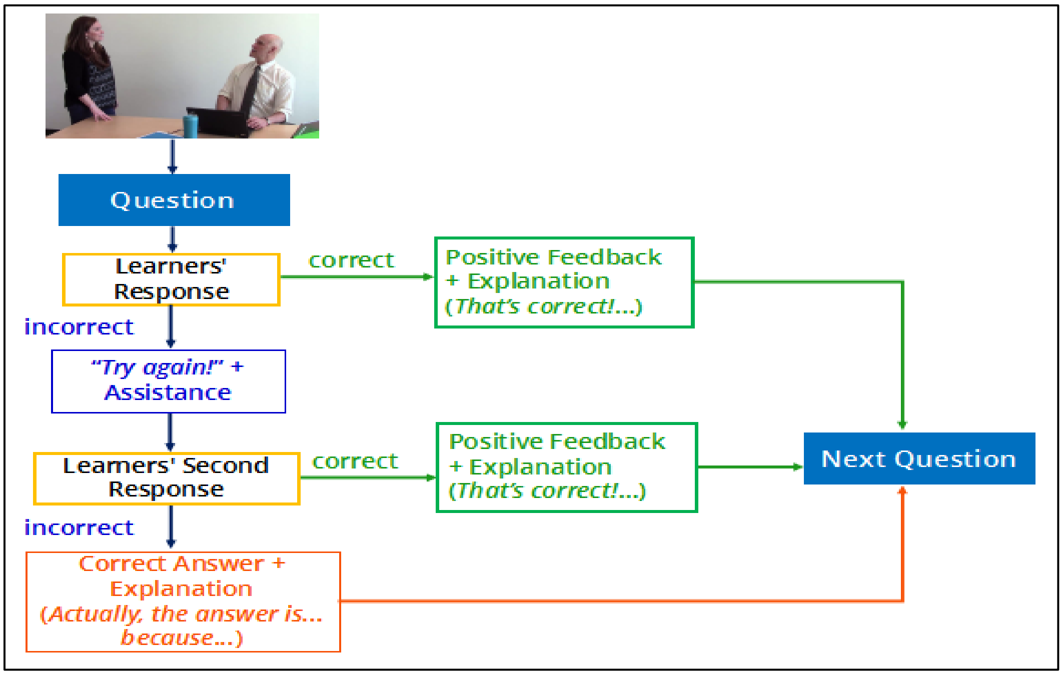

Question
Learners' Response
correct
Positive Feedback + Explanation (That's correct!...)
incorrect
"Try again!" + Assistance
Learners' Second Response
correct
Positive Feedback + Explanation (That's correct!...)
Next Question
incorrect
Correct Answer + Explanation (Actually, the answer is… because...)

Figure 7. Feedback loop design for items in Task One.

Task Two

The purpose of Task Two, the politeness judgment task, is to assess learners' ability to evaluate how polite a request needs to be worded in a particular situation based on the three contextual variables: relative social distance and power between interlocutors as well as imposition of the request.

Context statement. The context statement is carefully designed in order to include all three context variables in the intended constellation. The relationship between the interlocutors is described in a way that clearly outlines the relative status between interlocutors as well as the request that is to be made by one of the speakers. The context statement provides learners with a clear and concise picture of the situation, taking into account potential histories between speakers.

Prompt. The following generic prompt is used for each item type: Read the scenario and decide how polite the request should be. Then position the slider anywhere along the line from more polite to less polite.

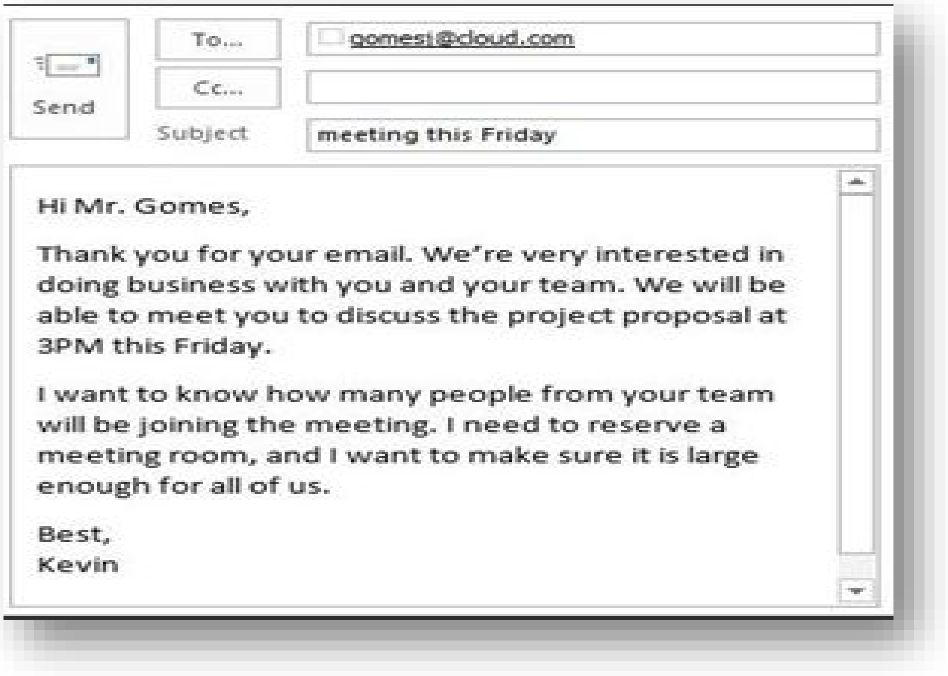
Feedback. Learners are provided with post hoc feedback in terms of how close their estimate is relative to the average of the responses provided by a particular reference group. In other words, in order to provide a benchmark or reference that would function as a point of comparison, responses would need to be collected from a carefully selected target group (e.g., L1 English speakers in a particular workplace in the United States). The mean of all responses provided by a particular reference group as well as the standard deviations around the mean could then be used as points of comparison to provide feedback.[3] For example, if a learner estimates the level of politeness two standard deviations below the mean of the reference group, they would receive feedback along the lines of (a) how close they were relative to the average of the reference/target group and (b) that the given reference group responses suggest the need for a more polite wording. Additionally, learners have the opportunity to listen to responses that have been identified as representative samples of appropriate requests made by L1 U.S.-English speakers in the given situation.

Task Three

The purpose of Task Three, the appropriateness judgment task, is to measure learners' ability to evaluate whether the request embedded in the interactions is pragmatically appropriate or not, and to elicit their ability to produce pragmatically appropriate language in both written and spoken forms. Each item includes a brief context statement, a prompt (an e-mail or a video), and a set of pragmatics judgment questions.[4]

Context statement. The context statement briefly describes the setting to introduce the reason for writing the e-mail, or the background of the interaction in the video. It is important to provide sufficient context information for students to evaluate whether the request made in the e-mail or in the video is appropriate or not. If a new character is introduced, his/her relationship with the interlocutor is described.

Prompt. There are two types of prompt in Task Three: an e-mail prompt for Part One, focusing on reading and writing skills; and a video prompt for Part Two, focusing on listening and speaking skills. The e-mail prompt resembles an actual e-mail. The structure of the e-mail follows what is taught in the learning module (Unit 3). The content is concise and written in a way that the appropriateness of the request is not ambiguous. A sample e-mail prompt is displayed in Figure 8.



gomesj@cloud.com

meeting this Friday

Hi Mr. Gomes,

Thank you for your email. We're very interested in doing business with you and your team. We will be able to meet you to discuss the project proposal at 3PM this Friday.

I want to know how many people from your team will be joining the meeting. I need to reserve a meeting room, and I want to make sure it is large enough for all of us.

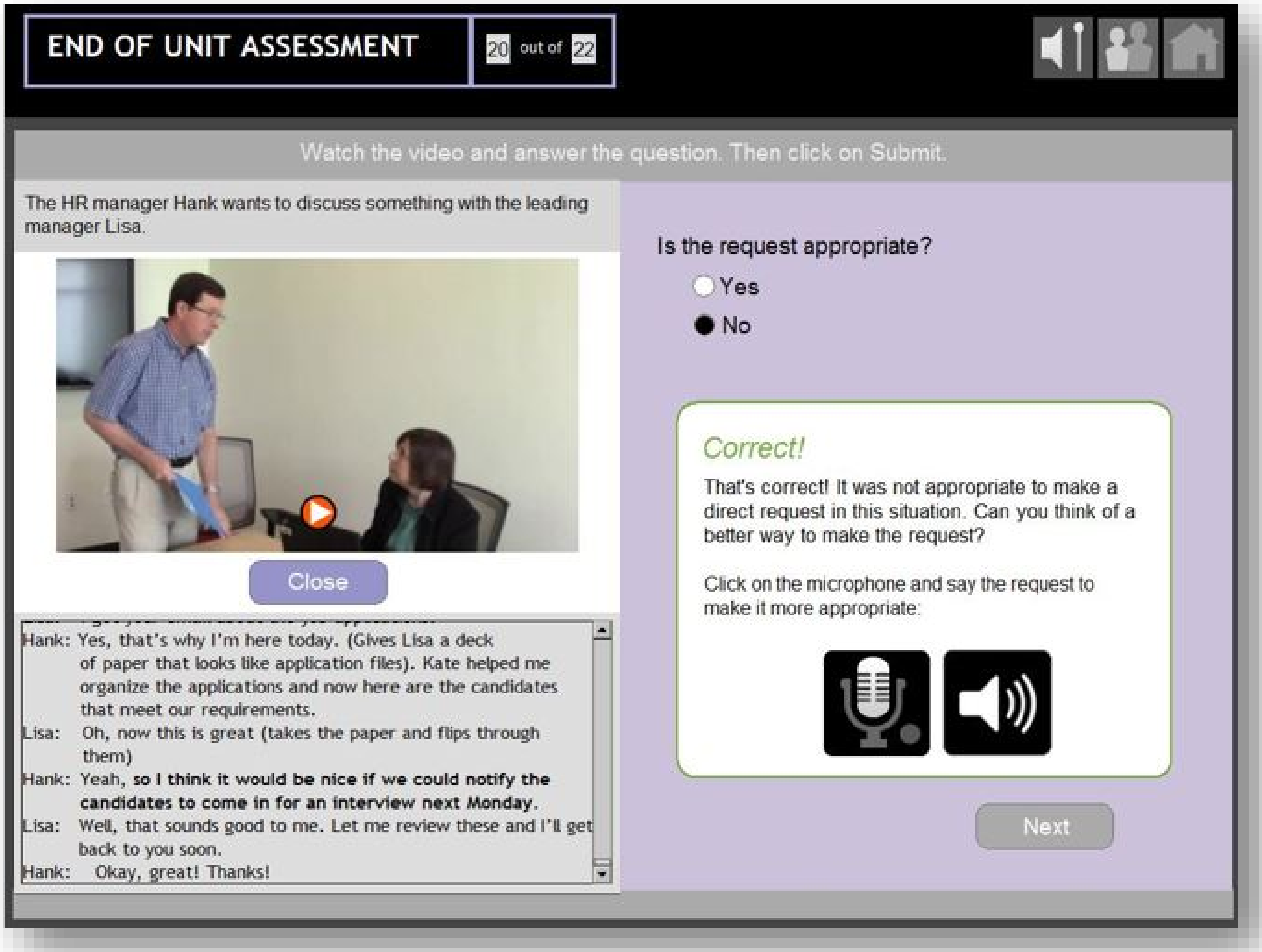
Best,
Kevin

Figure 8. Sample e-mail prompt from the appropriateness judgment task (Task Three).

In the case of audio-visual input, the video prompt generally follows the guidelines in Task One. However, it is important to remind the actors and actresses to use the “inappropriate” tone for inappropriate items (i.e., if the request is made in a rude way, the tone should reflect the rudeness).

Question sets. For both parts in Task Three, learners are presented with two questions (a) a pragmatic judgment dichotomous (yes/no) item asking whether the request is appropriate (i.e., “Is the request appropriate?”), and (b) a follow-up question asking for an appropriate alternative if the original request is not appropriate. That is, if the request embedded in the e-mail or the video as inappropriate, learners are prompted to write down (for e-mail items) or record (for video items) the appropriate alternatives (“What would you say in this situation?”). Additionally, they are asked to explain why or why not they evaluate the request as appropriate.

Feedback. Four types of feedback are implemented in Task Three. If the request is appropriate to the situation, and learners answer correctly, they will be given positive confirmation feedback (“That’s correct!”) with a brief explanation of why the request is appropriate for the purpose of learning reinforcement. However, if learners answer such a question incorrectly (i.e., consider an appropriate request to be inappropriate), they will be given a detailed explanation of why the request is, in fact, appropriate in this scenario. On the other hand, if the request is inappropriate to the situation, and learners answer correctly, they will be given an initial positive feedback (“That’s correct!”) and then be prompted to write down or record what would be appropriate to say. They are given the option to listen to their own recording and re-record once before submitting their response. After learners submit their response, they will see their own and a list of possible answers for them to compare their answers with. However, if learners cannot recognize the inappropriateness of the request, they will first be given scaffolded assistance that helps them understand why the request is, in fact, inappropriate. Then, learners will be asked to utilize the assistance and provide an alternative, appropriate response. After their response is submitted, a list of acceptable responses will be presented to them for comparison. Figure 9 shows the feedback learners would receive upon correctly identifying whether the request is appropriate to the situation. Moreover, it shows an optional scaffolding feature available to learners: the script to the video-delivered conversation between the two interlocutors.



Watch the video and answer the question. Then click on Submit.

The HR manager Hank wants to discuss something with the leading manager Lisa.

Hank: Yes, that's why I'm here today. (Gives Lisa a deck of paper that looks like application files). Kate helped me organize the applications and now here are the candidates that meet our requirements.
Lisa: Oh, now this is great (takes the paper and flips through them)
Hank: Yeah, **so I think it would be nice if we could notify the candidates to come in for an interview next Monday.**
Lisa: Well, that sounds good to me. Let me review these and I'll get back to you soon.
Hank: Okay, great! Thanks!

Is the request appropriate?

○ Yes
● No

Correct!

That's correct! It was not appropriate to make a direct request in this situation. Can you think of a better way to make the request?

Click on the microphone and say the request to make it more appropriate:

Figure 9. Feedback learners receive upon correctly identifying whether the request is appropriate to the situation.

Task Four

The purpose of Task Four is to evaluate a learner's ability to make requests in a multi-turn spoken conversation. Each item includes an instruction that provides the general context statement, and a number of prompt embedded in the dialogue structure.

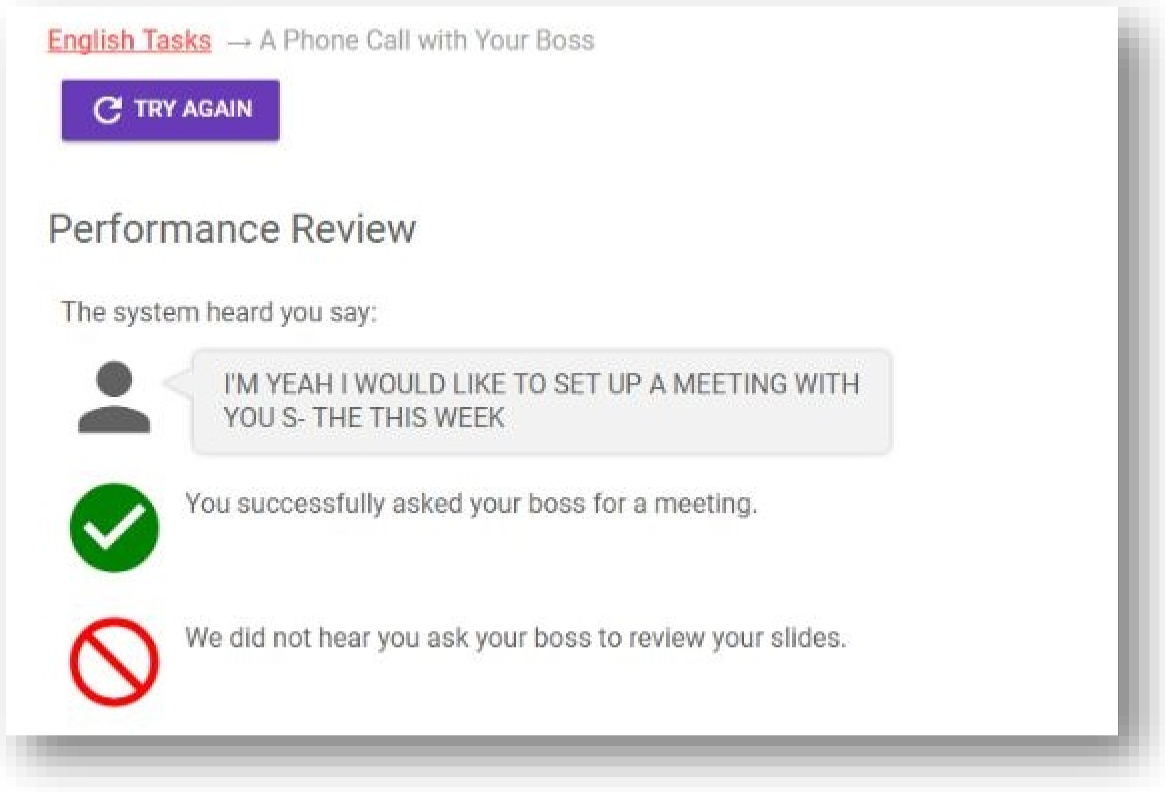
Instruction and context statement. In this task, the instructions provide a brief description of the context. They include the person they are calling as well as the purpose and goal of the phone call. If the person they are supposed to call is not a character from the learning module, then the relationship with the interlocutor will need to be described in more detail. The currently existing sample item features the following instructions: Imagine that you are calling your boss, Lisa Green.[5] Your goals are to (a) get her to agree to have a meeting with you and (b) ask her to review the presentation slides that you made so that you can discuss them at the meeting. Your schedule is free for the rest of the week so any time proposed by Lisa will work for you.

Prompts. The prompts in this item are embedded in the dialogue structure. A dialogue is designed to elicit at least one, ideally multiple requests. As shown in Table 2, the original request dialogue template is designed to elicit the request for a meeting in Turn 2 (T2) and the request for a review of the presentation slides prior to the meeting in Turn 4 (T4). However, given the variability in request behavior, the dialogue was then branched so the interlocutor (here Lisa Green) responds appropriately relative to when (i.e., in which turn) a learner makes the request(s). For example, a learner may make the requests subsequently in T2 and T4, while another learner may make both requests in a combined manner in T1. Thus, the automatic speech recognition needs to account for this type of variability and select a response that is appropriate based on when a user makes particular requests. For more detailed information on how to empirically develop a branched spoken dialogue tasks, please see Timpe-Laughlin et al. (2017).

Table 2. Request Dialogue Template

Turn	Interlocutor	Utterance
T1	Lisa Green	Hello?
	User	[greeting]
T2	Lisa Green	Hi, how's it going? What can I do for you?
	User	[(positive statement) + request for meeting]
T3	Lisa Green	Yeah, sure I'm available on Friday at 12. Does that work for you?
	User	[positive response]
T4	Lisa Green	Was there anything else you needed?
	User	[request to review slides]
T5	Lisa Green	Sure, no problem. Send them over.
	User	[expression of thanks]

Feedback. Learners get feedback with regard to (a) task completion and (b) appropriateness of the requests they made. The automatic speech recognition currently detects whether a user has made one-or-both of the requests. What a user says is transcribed in real time and provided to the user in the task-completion feedback following their completion of the call (see Figure 10). Additionally, learners will be provided with feedback based on the appropriateness of their requests (e.g., whether their request was too direct and thus slightly impolite or if they worded it appropriately). Finally, learners will be able to listen to L1 speaker samples as a form of feedback that has been found to be particularly effective in L2 pragmatics learning: modeling.



Performance Review

The system heard you say:

I'M YEAH I WOULD LIKE TO SET UP A MEETING WITH YOU S- THE THIS WEEK

✓ You successfully asked your boss for a meeting.

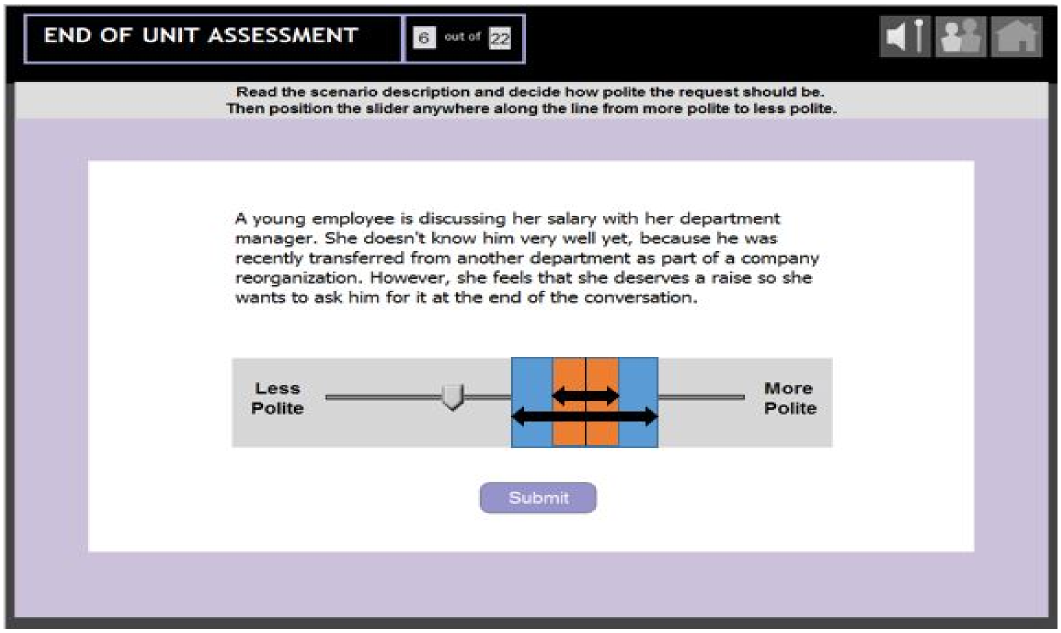
🚫 We did not hear you ask your boss to review your slides.

Figure 10. Task 4: Task-completion feedback.

Future Directions

While the prototype tasks of this assessment module have been designed and established, future steps need to be taken to hone scoring procedures and feedback delivery within the four task types. Given that the assessment module developed for *Words at Work* is primarily for formative assessment and self-evaluation purposes, formative feedback is more emphasized than summative scores. A particular focus in terms of next steps will be placed on the refinement of scoring procedures and feedback delivery. A particular goal will be to design a “performance report card” that matches the workplace theme of the learning tool and displays feedback vis-à-vis a learner’s relative strengths and weaknesses, while displaying correct and incorrect responses for each attempted task as well as recommendations for further development.

In order to further validate items and advance the feedback implementation and delivery, a number of steps will need to be taken in particular for Task Two, the politeness judgment task, and Task Four. For Task Two, feedback and scores will need to be established based on empirical evidence. Given that the item types under Task Two are prone to considerable variability, it is critical to carefully select and establish a clear reference group whose responses to the items can be used as a point of comparison. Thus, a large number of responses from a particular reference group such as L1 U.S.-English speakers will be collected to establish the continuous scale behind each item and set means that serve as a point of comparison to provide feedback. Then, the mean of all L1 speaker responses will be used to set a point estimate where the "appropriate answer" would be. The standard deviations around the mean can then be used to establish a range of acceptability. As shown in Figure 11, learners whose politeness ratings are within one standard deviation of the mean established on the basis of all reference group responses will receive full marks. If learners' ratings fall outside of one standard deviation but within two standard deviations of the mean, they will receive partial credit. If a learner's rating falls outside of two standard deviations, they will not receive any mark.



Read the scenario description and decide how polite the request should be.
Then position the slider anywhere along the line from more polite to less polite.

A young employee is discussing her salary with her department manager. She doesn't know him very well yet, because he was recently transferred from another department as part of a company reorganization. However, she feels that she deserves a raise so she wants to ask him for it at the end of the conversation.

Figure 11. Establishing a benchmark for scoring the evaluation items, using the confidence interval around the point estimate to establish a range of acceptability.

For Task Four, the multi-turn conversational-speaking task, feedback with regard to request appropriateness as well as L1 speaker samples are currently being implemented into the feedback landing page. For that endeavor, the L1 and L2 speaker responses collected for the item in Task 4 will be annotated for appropriateness and analyzed qualitatively to identify the most ommon request head acts (e.g., I was wondering if we could have a meeting vs. Can we have a meeting? vs. I request a meeting.). This information will then be used to train the language model to understand and gauge the directness and thus politeness of a request. Moreover, future research and development is planned to investigate the use of the *SpeechRater*^{SM} automated scoring service in combination with potential human ratings to provide more in-depth feedback about different speaking related phenomena in order to further inform and aid student learning.

In addition to refining the tasks in the request end-of-unit assessment, further steps need to be taken to further develop and operationalize the assessment modules across all four units implemented in *Words at Work*. Thus, it would be desirable to continue to expand the item bank and create additional self-assessment modules for the other units and speech acts using the ECD framework—a particular endeavor is currently under way for the small-talk unit.

References

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford, England: Oxford University Press.

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain.* New York, NY: David McKay Company.

Clyne, M. (1994). *Inter-cultural communication at work: Cultural values in discourse*. Cambridge, England: Cambridge University Press.

Cohen, A. D. (2008). Teaching and assessing L2 pragmatics: What can we expect from learners? *Language Teaching, 41*, 213–235. https://doi.org/10.1017/S0261444807004880

Crystal, D. (Ed.). (1997). *The Cambridge encyclopedia of language* (2nd ed.). New York, NY: Cambridge University Press.

Grabowski, K. C. (2009). *Investigating the construct validity of a test designed to measure grammatical and pragmatic knowledge in the context of speaking*. Unpublished doctoral dissertation, Teachers College, Columbia University, New York, NY.

Grabowski, K. C. (2013). Investigating the construct validity of a role-play test designed to measure grammatical and pragmatic knowledge at multiple proficiency levels. In S. Ross & G. Kasper (Eds.), *Assessing second language pragmatics* (pp. 149–171). New York, NY: Palgrave Macmillan. https://doi.org/10.1057/9781137003522_6

Grabowski, K. C. (2016) Assessing pragmatic competence. In D. Tsagari & J. Banaerjee (Eds.), *Handbook of applied linguistics: Vol. 12. Handbook of second language assessment* (pp. 165–180). Boston, MA: De Gruyter Mouton.

Leech, G. N. (1983). *Principles of pragmatics*. London, England: Longman.

Levinson, S. C. (1983). *Pragmatics.* Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511813313

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design* (Research Report No. RR-03-16). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

Mislevy, R. J., & Riconscente, M. M. (2005). *Evidence-centered design: Layers, structures, and terminology* (Report No. 9). Menlo Park, CA: SRI International.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (1999). *Evidence-centered assessment design.* Princeton, NJ: Educational Testing Service.

Purpura, J. E. (2004). *Assessing grammar*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511733086

Purpura, J. E. (2016). Second and foreign language assessment. *The Modern Language Journal, 100*(S1), 190–208. https://doi.org/10.1111/modl.12308

Ramanarayanan, V., Suendermann-Oeft, D., Lange, P., Mundkowsky, R., Ivanov, A., Yu, Z., Qian, Y., & Evanini, K. (2017). Assembling the jigsaw: How multiple open standards are synergistically combined in the HALEF multimodal dialog system. In D. Dahl (Ed.) *Multimodal interaction with W3C standards: Toward natural user interfaces to everything* (pp. 295–310). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-42816-1_13

Roever, C. (2009). Teaching and testing pragmatics. In M. H. Long & C. J. Doughty (Eds.), *The handbook of language and teaching* (pp. 560–576). West Sussex, UK: Blackwell Publishing Ltd. https://doi.org/10.1002/9781444315783.ch29

Taguchi, N., & Sykes, J. (Eds.). (2013). *Technology in interlanguage pragmatics research and teaching*. Philadelphia, PA: John Benjamins. https://doi.org/10.1075/lllt.36

Timpe-Laughlin, V. (2016). *Learning and development of second and foreign language pragmatics as a higher-order language skill: A brief overview of relevant theories* (Research Report No. RR-16-35). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12124

Timpe-Laughlin, V., Evanini, K., Green, A., Blood, I., Dombi, J., & Ramanarayanan, V. (2017). Designing interactive, automated dialogues for L2 pragmatics learning. In V. Petukhova & Y. Tian (Eds.), *Proceedings of the 21st workshop on the semantics and pragmatics of dialogue* (pp. 143–152). Retrieved from http://www.saardial.uni-saarland.de/wordpress/wp-content/uploads/SemDial2017SaarDial_proceedings.pdf

Timpe Laughlin, V., Wain, J., & Schmidgall, J. (2015). *Defining and operationalizing the construct of pragmatic competence: Review and recommendations* (Research Report No.

RR-15-06). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12053

Washburn, G. (2001). Using situation comedies for pragmatic language teaching and learning. *TESOL Journal, 10*, 21–26.

Yule, G. (1996). *Pragmatics*. Oxford, England: Oxford University Press.

Zieky, M. J. (2014). An introduction to the use of evidence-centered design in test development. *Psicología Educativa, 20*, 79–84.

Notes

[1] Unit 3 deals with requests in emails and general workplace requests.

[2] Note that the name *John* is used as a generic placeholder. The name varies depending on the characters that are featured in the video.

[3] Please note that further research is necessary to develop this feature.

[4] In order to allow students more opportunities to demonstrate their writing and speaking skills, there are more inappropriate items than appropriate ones.

[5] Note that Lisa Green is the boss character in the learning module.