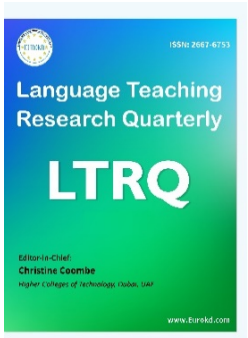

Language Teaching Research Quarterly

2023, Vol. 37, 188–203

An AI Generated Test of Pragmatic Competence and Connected Speech

Stefan O'Grady

International Education Institute, University of St Andrews, United Kingdom

Received *04 June 2023* ***Accepted*** *14 October 2023*

Abstract

Language testing is witnessing increasing interest in the potential for AI to support test development and validation. To date, published research involving AI in language testing has typically been conducted in the context of high-stakes proficiency tests and the potential for this technology to support local language testing is under-researched. The current study seeks to address this gap by reporting on the piloting of an AI generated language test in the context of a university in the UK. The focus of the paper was selected to explore two key areas in the work of J.D. Brown, namely pragmatics and connected speech. In the study, international students with English as a second language completed an AI generated test of pragmatics created from a transcript of spontaneous interaction, and a test-taking strategy questionnaire. Results demonstrated that the test did not provide a sufficiently reliable measurement of test takers' pragmatic competence, although reliability estimates did improve with the removal of problematic items. Implications for language test development involving AI are discussed.

Keywords: *Pragmatic Competence, Connected Speech, Artificial Intelligence, Language Testing*

Introduction

The current study seeks to combine two consistent areas of focus in the work of J. D. Brown: second language (L2) pragmatics and connected speech. The motivation for this endeavour reflects a common theme in Brown's work of investigating approaches to the design of language assessments that represent authentic language use and language processing effectively; "in many respects, language testing is about the careful design of tests so they will prove reliable and valid" (Brown, 2008, p. 226). The specific aim of the current study is to move towards authentic language use in L2 listening assessments (Brown & Trace, 2018) by outlining the development and evaluation of a prototype listening test of pragmatic competence. To this end, the study builds on increasing interest in artificial intelligence (AI)

* Corresponding author.
E-mail address: so59@st-andrews.ac.uk
https://doi.org/10.32038/ltrq.2023.37.10

in language testing (Mizumoto & Eguchi, 2023; O'Sullivan et al., 2023), and evaluates the potential for AI to support test development. The study investigates a test involving items that were generated by ChatGPT and represents an effort to determine how AI tools might be used in local contexts that lack the resources of large testing organisations but nevertheless seek to gather reliable information about individual's L2 competence with this technology.

Background

Pragmatics

Pragmatics refers to the rules that govern the linguistic choices we make in communication based on our understanding of contextual factors of the situation like social distance and power differentials between individuals (Brown, 2016). Brown contributed to pioneering work on the assessment of L2 pragmatic competence in Hudson et al. (1992, 1995). This was work with an important focus that represented "the first effort by language testers to systematically develop and examine the effectiveness of tests of pragmatics ability" (Brown, 2008, p. 225). Since then, interest in assessing pragmatics has grown substantially (Ishihara & Cohen, 2021; Roever, 2011). As a communicative construct, pragmatic competence is now widely assessed in modern language tests and is a core component of communicative competence outlined in the Common European Framework of Reference for Languages (Council of Europe, 2001, 2018). Brown and Ahn (2011, p.199) explain that pragmatic competence engages four core variables, namely "(a) differences in the functions of speech acts (e.g., requests, apologies, refusals, etc.), (b) differences in the relative power of speakers and hearers, (c) differences in the social distance of speakers and hearers, and (d) differences in the degree of imposition required or perceived in a given speech act." Assessment tasks seeking to measure pragmatic competence must therefore necessarily reflect these variables and require test candidates to demonstrate awareness and sensitivity to these aspects of relevant communicative situations.

In the literature, a range of tasks have conventionally been applied to measure pragmatic competence (Roever, 2011). These include discourse completion tasks in which the learner is presented with a partial recording or transcription of an interaction and is required to complete the interaction by providing a written, spoken or selected response to a question or comment in the discourse; self-assessment tasks in which the learner predicts their ability to navigate a hypothesised situation in the L2; role play tasks; and role play self-assessment tasks involving an evaluation of a recording of the learner completing the task. Crucially, each task is associated with a test method effect, e.g. overestimation of ability during self-evaluation or differences in rater consistency and severity when scoring responses, that may distort interpretation of results in a way that test scores represent more than the construct of interest (Brown & Ahn, 2011). To be sure, developers of assessments of pragmatic competence would do well to acknowledge the inherent limitations of each format and include a variety of task types to mitigate this.

Connected Speech

Contemporary L2 listening assessments are often developed from scripts rather than authentic language use (O'Grady, 2021). The immediate advantages of this approach are often clear to individuals with a responsibility for developing listening assessments. Scripting allows

assessment developers to create a neat and tightly organised sound file that fosters production of comprehension questions presented at equal intervals with clear and unambiguous answers, often using a selected response format such as multiple-choice that can be objectively scored (Wagner & Ockey, 2018). However, in the long term the consequences of assessing L2 competence in this way are likely to be harmful; the underlying flaw in this approach is that the organisational control comes at the cost of authenticity (Wagner, 2018).

Authentic patterns of spoken interaction involve connected speech (Brown, 2006). Connected speech refers to the ways in which pronunciation often anticipates and is affected by adjoining words in an utterance and may be distinguished from speech in which words are produced in citation forms. For example, the following exclamation may be produced using citation forms or connected forms:

A) I asked you not to do that!

B) I(y)asch'e no' t' do tha'!

Whereas the former might be reserved for forms of sarcasm or chiding, e.g. an exasperated parent or teacher explaining something to a misbehaving child, the latter may represent the more conventional and commonplace form. While a stigma has historically associated connected speech with laziness, Brown and Kondo-Brown (2006) convincingly outline the case for its inclusion in L2 listening pedagogical material, explaining that connected speech is natural speech, something students need to know and in fact often enjoy learning about. In a recent study examining the use of connected speech in dictation activities, Brown and Trace (2018, p. 62) conclude that "a more accurate and complete representation of listening ability must include a valid representation of real-world spoken language, and therefore, connected speech must receive its due in the language assessment community." According to this line of reasoning, sound files developed to assess L2 pragmatic competence should include characteristics of connected speech relating to "word stress, sentence stress and timing, reduction, citation and weak forms of words, elision, intrusion, assimilation, juncture, and contraction" (Brown, 2006, p. 15).

Crucially, for sound files to be regarded as genuinely authentic, assessment developers must focus on both connected speech and spoken grammar. Spoken grammar refers to patterns of grammar that would be uncommon in formal written language but are nevertheless systematic in speech, such as fillers (y'know), tag questions (innit?) or situational ellipsis ("A. How was the film? B. (I) Didn't go in the end"; Brown & Crowther, 2022; Carter & McCarthy, 2006). An interesting parallel can be drawn regarding distinctions between "grammar" and "spoken grammar" on the one hand, and "speech" and "connected speech" on the other. The spoken and connected forms have traditionally been regarded as deviant, while the written and citation forms are perceived as the standard. A clear consequence of this, language assessments are developed according to these standards and test candidates are certified as having a particular level of ability in a second language on the basis of a test featuring forms that bear limited resemblance to authentic communication. For language test scores to extrapolate to authentic social interaction in the L2, it stands to reason that the test must feature such forms. Furthermore, in the context of the aims of the current study, increasing automation of language test development underscores an urgent need to maintain a high level of representation of the

complexities of authentic human interaction in L2 assessments. Though there may be some immediate resistance as a result of the introduction of authentic language in listening tests in contexts that have not traditionally tested such forms, e.g. due to low test scores stemming from a traditional focus on citation and written forms in classrooms, this should eventually be mitigated by the need to base instruction on authentic language.

AI in Language Testing

Interest in the role of AI in language testing has grown rapidly in recent years. Recent studies have investigated the use of AI in reading test generation (Attali et al., 2022), automated scoring (Mizumoto & Eguchi, 2023; O'Sullivan et al., 2023), and the evaluation of the complexity of writing prompts (Khademi, 2023). However, to date the area remains under-researched and there are no published studies investigating the potential for AI to generate test items from samples of authentic spoken language. The current study seeks to contribute to the incipient body of literature by evaluating the use of AI to develop test items for a listening task in a local language test. The development of local language tests has traditionally involved groups of individuals creating test items from specifications and piloting the items with members of the test taking population (Dimova et al., 2020). This is a resource intensive endeavour that takes place within strict institutional constraints. The reality is that many test developers producing tests for foreign language programs or academic language preparation courses for university study often juggle teaching and administrative responsibilities with the result that a limited number of test items may be produced in a given time. Test security is thus a major concern. A solution that has been explored in the literature is to reduce the workload by developing test items with generative AI (Settles et al., 2020; Yunjiu et al., 2022). Test items created with AI are becoming increasingly common in language tests, such as the Duolingo English Test (Cardwell et al., 2022). However, research investigating the potential for AI generated items to measure specific constructs of interest to language testers, such as pragmatic competence as outlined by Brown (2016), is limited. The current study seeks to address this gap.

Problem Statement

Tests of L2 ability should generate information about test candidates' pragmatic competence as a core component of their communicative competence. For test scores to extrapolate, the demonstration of pragmatic competence should be elicited in response to samples of authentic language use. Current approaches to language assessment development favour organisational control over language authenticity with the implication that test scores do not adequately reflect test candidates' pragmatic competence with authentic language. Furthermore, the process of developing test items is resource intensive and generative AI may provide a valuable means of support. The current paper therefore investigates an approach to develop tests of pragmatic competence using authentic speech with AI generated test items and evaluates the compromises in test design that may need to be made when utilising this technology.

Procedures

Developing the Sound File

Structured improvisation has been promoted as a way to create sound files that involve authentic speech but simultaneously allow the test developer to retain some organisational control of the content (Clark, 2014). Applying this approach, a team of developers create an outline of a communicative situation that test candidates are likely to navigate in the L2. This outline is provided to voice actors (individuals responsible for recording the speech) who 'follow the basic outline of the text while still composing and uttering (virtually) simultaneously' (Wagner & Ockey, 2018, p. 21). The outcome of this process is a sound file that contains spontaneous and authentic language use and is therefore considered more representative of real-world spoken language. However, because a broad outline of the event has been followed, the test developer may exert some control over the sound file to meet test specifications. The sound file may then be applied to develop items that measure test candidates' pragmatic competence in processing authentic speech.

Following the procedures outlined above, a sound file for a test of pragmatic competence for tertiary level English medium education in the UK was developed. Firstly, a communicative situation that would be common in university study in an English medium environment was identified by the researcher by reflecting on common experiences on the university campus. The situation was a student returning books at a university library. The recording was created by one male speaker of British English and one of Canadian English. The Canadian speaker assumed the role of the student and was told that he was returning some books. The British speaker assumed the role of the librarian and began the recording with the knowledge that the Canadian was returning some books but also with the foreknowledge that one of the books was late. This extra information was important to elicit anticipated speech acts such as a warning, apology, or a request. For the sound file to appear authentic, it was also important for the speech acts to be natural and unrehearsed and withholding this information from one of the speakers was anticipated to facilitate this. The interaction was recorded as an MP3 on a laptop computer and the two voice actors spoke for approximately two minutes.

The speech was transcribed, and phonological analysis was carried out to identify examples of features of connected speech (Roach, 1998; see Appendix 1). The analysis identified several features of connected speech:

- Reduction: a process by which phonemes are diminished or excluded to facilitate smooth pronunciation.

lots o' people (of pronounced as a schwa /ə/)

- Citation and weak forms of words: in the citation form, all phonemes in a word are produced whereas in connected speech some may not. In the following example both the weak form and citation form of "last" are produced, reflecting the speaker's emphasis and perceived need for clarity about the date.

*this las' book erm was due **last week***

- Elision: the loss of vowels or consonants that would be pronounced in the citation form of a word.

Yes nex' please

- Intrusion: the insertion of a sound to facilitate transition between syllables or words. In the following example, *all* is pronounced /rɔːl/ after the preceding /r/ in *were*:

I though' they were'all due back on the same day

- Juncture: the transition between phonemes at word boundaries or within a single word. In the following example, there is no clear break between the words "at" and "all":

there's no rush at'all

Contraction: a process whereby a word, commonly the auxiliary verb have, the participle not, the verb to be, and modal verbs will and would, are not pronounced in full and are attached to the preceding word.

Righ' I'll jus' log you into the system

In addition, the recording contained various features of spontaneous speech and spoken grammar relating to pausing and hesitation (*Er yes I' like t' return these er books erm I think they're due today I'm not sure*), repetition (*tha' tha' makes sense*), false starts (*erm ne relatively new books like this*), and overlap between speakers (overlapping speech is presented directly below the interrupted line in square brackets):

1. *Ok grea' sorry 'bout that I din' realise*

2. *[Tha's ok]*

1. *so wha's the normal erm loan perio'fe books*

Conventional approaches to developing tests of pragmatics

The conventional approach to develop a task of pragmatic competence involving the transcribed sound file would target the key speech acts in the transcript (Brown & Ahn, 2011; Roever, 2011). For example, in the following exchange, the student's response could be removed from the sound file and replaced with a multiple-choice item that included the original response as the key among a series of alternative response options. This item would provide information about the test taker's ability to apologise in a way that would be deemed appropriate in the context.

Test candidates hear the following and select the most appropriate response:

1. *Ah 'cause I though' they were all due back on the same day er are you sure abou' that*

1. *[Yes er] the computer seems t' think so anyway erm I nee t' tell you there's a fine involved'of two pounds la' fee*

2. *Ok erm well I don' have money right now can I bring this afternoon*

1. *[Yes] of course th tha tha's fine there's no rush at'all*

Select the most appropriate response.

A. Ok great sorry about that I didn't realise.

B. Yes, I'm so so sorry, this will never happen again.

C. Indeed, my sincerest apologies, how can I make it up to you?

In the task, the test taker listens to the majority of the recording to establish important contextual information such as status differentials and selects the most appropriate response given their understanding of the communicative context. Such selected response tasks have the advantage of quick and objective scoring, however as Brown (2016) points out, are only capable of assessing passive knowledge and not the ability to demonstrate pragmatic competence independently. For this reason, a constructed response in which the test taker writes or speaks their answer at the end of the recording may be preferable. Advances in computer-based assessment have made the use of spoken responses to this kind of pragmatics test increasingly possible. A further advantage of designing these assessments to be delivered by a computer is that the test taker's responses may be timed, and extra points awarded for speed, reflecting processing demands in live interaction (Taguchi, 2008). The computer mediated task may also draw upon adaptive testing formats, whereby specific items are identified to enhance the measurement on the basis of accuracy and speed of response (Kaya et al., 2022).

In classroom based formative assessment, the format may also lend itself to a self-assessment task whereby test takers predict their ability to participate in the situation:

In this situation, my response would probably be…

Totally unacceptable 1 - 2 - 3 - 4 - 5 Very effective

This kind of task would be helpful to teachers seeking to develop their students' ability to self-diagnose and encourage more autonomy among their learners (Tai et al., 2018). As described previously, the test method effects associated with each response format entail that a combination of methods may be necessary to generate a reliable profile of a test candidate's pragmatic competence.

Generating the Test with AI

Having established the traditional means of development of tests of pragmatic competence, the following section outlines the procedure for developing the test featuring AI generated content. ChatGPT, an AI chatbot produced by OpenAI (https://openai.com/blog/chatgpt), was used to create the test of pragmatic competence and an answer key using the transcript of the sound file. The full instructions given to the AI drew upon the definition of the construct outlined by Brown and Ahn (2011, p.199).

Instructions to ChatGPT:
Create a listening test of pragmatic competence from this transcript with answer key. The test should assess (a) differences in the functions of speech acts (e.g., requests, apologies, refusals, etc.), (b) differences in the relative power of speakers and hearers, (c) differences in the social distance of speakers and hearers, and (d) differences in the degree of imposition required or perceived in a given speech act:
Librarian: Yes next please
Student: Er yes I'd like to return these er books erm I think they're due today I'm not sure
Librarian: Ok that's great could I ask for your university ID number please
Student: Yes its er nine one six er two zero five one eight two
Librarian: Right I'll just log you into the system here take the books thank you erm that seems to be fine yeah oh erm this last book erm was due last week
Student: Really?
Librarian: Yeah er it says that you checked it out on the twenty seventh and it's a one week book so it should have been returned on the fourth
Student: Ah cause I thought they were all due back on the same day er are you sure about that
Librarian: Yes er the computer seems to think so anyway I need to tell you there's a fine involved of two pounds late fee
Student: Ok erm well I don't have money right now can I bring this afternoon
Librarian: Yes of course that's fine there's no rush at all
Student: Ok great sorry about that I didn't realise
Librarian: That's ok
Student: so what's the normal erm loan period for books
Librarian: well erm it depends on the depends on the book erm re relatively new books tend to be one week only
Student: ok I suppose because lots of people wanna borrow that book right ok that that makes sense ok

The model was asked to regenerate the response a total of ten times before an acceptable series of test items was produced (see next section). This was typically due to the generation of items that did not clearly measure pragmatic competence (e.g. "How much is the fine for returning the last book late? A. One pound B. Two pounds C. Three pounds D. Four pounds") or because the items involved reading rather than listening ("so what's the normal erm loan period for books" a. Request b. Offer c. Refusal d. Statement of fact").

ChatGPT created test of pragmatic competence:

Listen to the conversation between a student and a librarian and answer the following questions.

1. What is the student's speech act in the first turn? A. Request B. Apology C. Refusal D. Assertion Answer: A. Request

2. What is the librarian's power in the conversation? A. High power B. Low power C. Equal power D. Power is not relevant in this conversation Answer: A. High power
3. What is the social distance between the student and the librarian? A. Close social distance B. Moderate social distance C. Far social distance D. Social distance is not relevant in this conversation Answer: B. Moderate social distance
4. What speech act does the librarian perform when she* informs the student about the late fee? A. Request B. Apology C. Refusal D. Warning Answer: D. Warning
5. What speech act does the student perform when he asks if he can bring the money in the afternoon? A. Request B. Apology C. Refusal D. Offer Answer: A. Request
6. What is the degree of imposition in the librarian's speech act? A. Low imposition B. Moderate imposition C. High imposition D. Imposition is not relevant in this conversation Answer: B. Moderate imposition
7. What is the function of the student's speech act in the last turn? A. Request B. Apology C. Refusal D. Assertion Answer: A. Request**
8. What is the function of the librarian's speech act in the last turn? A. Answering B. Clarifying C. Agreeing D. Disagreeing Answer: B. Clarifying

*changed to "he" in the pilot test ** changed to assertion

Once the model had produced a series of satisfactory items, the proposed answer key was evaluated by the researcher and an English teacher at the university who took the test and supplied responses. Overlap with the proposed answer key was observed with the exception of question 7, which identified the student's final contribution ("ok I suppose because lots of people wanna borrow that book right ok that that makes sense ok") as a request. This item seemed to ask about the previous contribution and the answer key was updated. The potential for references to first and last turns and terminology such as "speech act" to cause confusion was also noted by both the researcher and the teacher. The researcher also noted a series of potentially ambiguous terms in the items e.g. "librarian's power" and "social distance", though the decision was made to retain them to minimize the number of alterations made to the AI content. However, a final change was required because the model referred to the librarian as "she" in the fourth item, reflecting a possible tendency of the AI to ascribe sex to specific professions that should be on the radar of language educators that are considering the use of this technology for pedagogical purposes. Based on the items the model produced, the AI generated test questions requiring the test taker to demonstrate awareness of features of the social situation, speaker intent, power differentials, social distance, and degree of imposition. As the AI generated task does not require test takers to demonstrate pragmatic competence independently but rather to show sensitivity and awareness of the pragmatic aspects of the social situation, it may be appropriate to regard the test more precisely as a measurement of meta-pragmatic awareness (Ishihara & Cohen, 2021).

Evaluating the Task

The process of assessment task evaluation involves both short-term and long-term processes of inquiry. In the short-term, there are two general areas of focus. Broadly, these relate to the

technical qualities of task scores (Brown, 2014) and the cognitive processes test takers engage in to complete the task (Field, 2013). On the former, Brown has made important contributions in raising awareness about the role of statistical analysis among language testers, not least in the JALT Testing and Evaluation SIG newsletter (https://hosted.jalt.org/test/pub.htm), which has provided accessible answers to complex statistical questions to a generation of language professionals. Following Brown and Hudson (2002), establishing the technical qualities of the task may involve test takers completing the task and analysis of responses using statistical techniques from classical test theory or item response theory.

Evaluating the task further requires information gathered through questionnaires, focus groups or introspective interviews to determine the cognitive processes test takers engage in (Field, 2013). This aspect of the evaluation focusses on the extent to which the task requires a) processes that would be employed in the target situation or is influenced by test characteristics that introduce construct irrelevant variance and b) a range of processes rather than a small subset that would indicate construct underrepresentation (Messick, 1989). In listening assessment research, there is a precedent to utilise these kinds of methods to venture into the mind of the listener and establish what Field (2011) refers to as cognitive validity. In the assessment of pragmatics, research studies have also investigated the cognitive validity of tasks with these methods (Youn & Bi, 2019). Increasingly, findings from such studies are supported with new technologies such as eye-tracking (Holzknecht et al., 2021) and neuro-imaging (Aryadoust et al., 2022), though the cost and technical expertise required by such methods means this is probably beyond the remit of language teachers developing local assessments at present.

To evaluate the current AI generated task, the test was sent to 16 international students with English as a second language enrolled at a university in the UK who completed the test and the survey online using Qualtrics (https://www.qualtrics.com/uk/). Ages ranged from 22 to 64 (mean age = 32.2) and the students came from a variety of L1 backgrounds including Burmese, Chinese (8), Japanese (3), Russian, Slovak, Vietnamese, and Arabic. The minimum entry requirements for the university are B2 level proficiency in English (Council of Europe, 2001). Each item on the test was weighted equally as one point per item. Item facility and point-biserial correlations are reported to examine individual item characteristics and information about overall test consistency is established using Cronbach's alpha. The mean score on the test was 3.88 and the standard deviation was 1.32. These figures imply that the test takers found the test quite difficult. Table 1 presents the item facility and point biserial values for each item. The item facility statistics range from .19 to .94 indicating that the test featured a variety of difficult and simple items. Evaluating the point-biserial correlation statistics, items 2 and 5 do not correlate with overall scores and may require revision or deletion. The remaining items have a moderate positive correlation. Cronbach's alpha value was .07 and the standard error of measurement was 1.27, indicating that the test did not generate a convincingly reliable measurement. However, upon removing items 2 and 5, the average score remained comparable with an average of 50% accuracy in responses (mean = 3.19, SD = 1.38), whereas Cronbach's alpha value was .49 (SEM = 1.02); though this value indicates poor reliability, this is a substantial increase (Brown & Ahn, 2011).

To provide evidence of the test takers' engagement with the task, the study adapted a test taking strategies self-report questionnaire developed by Low and Aryadoust (2021). Participants completed the questionnaire online by confirming their level of agreement with each statement on a scale of 0 to 100. The results are presented in Table 2. The mean values are presented as indicative of the participants' engagement with the task however, these patterns should be interpreted in the light of the frequently large standard deviation values, which indicate that there was often substantial disagreement between the participants. The results indicate that participants approached the task strategically by prioritising contents of the items (1), and that the items were accessible (6) and provided a contextualising focus (2 & 5). The task appears to encourage test takers to focus on main ideas and to make inferences about the sound file (8 & 9), implying that the test is measuring higher order listening processes (O'Grady, 2021). It was common for students to draw upon background knowledge to answer the questions (21), although this may be an indication that it was possible to accurately respond to the items without necessarily comprehending the sound file, this is unlikely given the nature of the questions (11, 12, 15) and may well refer to participants drawing upon background knowledge about the context to answer the questions (3). Finally, few students reported taking notes (14).

Table 1

Item Statistics

Item	Item Facility	Point-biserial correlation
1	.50	.37
2	.25	.05
3	.94	.55
4	.19	.40
5	.44	.08
6	.63	.59
7	.50	.55
8	.44	.36

Table 2

Survey Responses

Statement		Mean	SD
1	I paid more attention to some words or phrases in the questions and options.	74.06	21.19
2	I predicted a general idea of the conversation based on the questions and options.	68.20	24.69
3	I tried to answer the questions based on my previous knowledge about this topic.	49.20	35.71
4	I guessed the meaning of some unfamiliar words or phrases from the questions and/or options.	57.13	26.17
5	I used the questions as an outline of the conversation.	66.40	32.27
6	I could easily understand the questions and the options.	70.25	21.91
7	I mainly listened for the details required by the questions.	65.75	33.52
8	I mainly listened for the main ideas to answer specific questions.	74.73	20.96
9	I made inferences based on the conversation.	79.69	14.56
10	The conversation was different from what I had guessed during question preview.	45.79	34.47
11	It was easy to match the keywords from the conversation and the questions.	56.63	22.82
12	It was easy to match the keywords from the conversation and the options.	51.88	22.58
13	I used clues from other questions to answer the question at hand.	57.14	27.41
14	I took notes from the conversation and referred to them to answer some questions.	21.00	28.16
15	I answered the questions based on common sense.	47.53	32.11

16	I tried to determine which questions to respond based on the progress of the conversation.	67.13	26.36
17	In some cases, I tried to stop listening and focus on item-reading.	54.40	30.65
18	I had to listen to the conversation and read the questions at the same time.	47.13	35.23
19	Overall, I did not have to spend a long time to read and understand the questions/options.	56.38	28.72
20	The questions helped me understand the conversation better.	64.50	23.17
21	I did not use my background knowledge to answer the questions.	28.20	31.88
22	I am confident about my answers.	58.44	24.64

Discussion

The study investigated the potential to use AI to generate test items from a transcript of authentic speech to measure L2 pragmatic competence. The AI generated test was shown to lack sufficient reliability as a measurement of the construct. While limitations in sample size and the number of items impact on the reliability of the measurement (Brown, 2002), it is clear that the AI generated items did not elicit a consistent, systematic pattern of responses from the participants. However, upon removing two problematic items, Cronbach's alpha increased to a level that suggests with more participants the test may demonstrate sufficient internal consistency to measure pragmatic competence. In addition, the results of the survey indicate that the processes involved in completing the task were largely appropriate and relevant, though there was variation between participants, implying that information regarding their interpretation of the survey questions would also have been valuable. With careful amendments to distractors, the test may provide a reliable measurement to be included in a suite of tasks involving discourse completion and self-assessment to collect information about test candidates' levels of pragmatic competence.

In the long term, as language assessment developers we are interested in the washback of introducing new assessment tasks into an educational environment (Cheng & Sultana, 2021), and the potential for the test to predict future behaviour (Isaacs et al., 2023). Though washback is often difficult to predict, it can be assumed that a focus on pragmatics and connected speech might encourage students to consider the relationship between authentic language use and social context and could hence be beneficial, provided that the necessary resources are available to teach and learn these skills (Green, 2020). However, it is unclear how test candidates and other stakeholder groups would respond to the knowledge that the test questions were created by AI and whether this would impact on their approaches to test preparation or their willingness to accept the test as appropriate for its intended use. It is also uncertain whether the AI operationalisation predicts pragmatic competence in social situations in the L2 in a similar way as conventional assessments of the construct (Brown & Ahn, 2011). These are important concerns that might be investigated through lesson observations, and follow-up surveys and interviews that seek to determine the role that pragmatic competence as operationalised by the AI plays in the test takers' experiences in the L2.

The current study underscores the emerging potential to measure pragmatic competence with connected speech using test items generated by AI. This is an important focus as test developers seek new ways to increase the representativeness of assessment tasks and the utility of test scores while making use of the possibilities afforded by new technologies. Overall, the

test was shown to lack reliability but did seem to engage relevant processes. Ultimately, AI may support test development but careful evaluation of items and item statistics and considered modifications by language test developers remains necessary.

Conclusion

The present study critically evaluates an AI generated L2 listening assessment of two key areas of interest in the work of J.D. Brown i.e., L2 pragmatics and connected speech. It is important to acknowledge several limitations in the research. Primarily, the sample size constrained the analysis of the results to rudimentary statistical methods and no doubt impacted on the reliability of the assessment. It would have been informative to analyse responses with item response theory models and a larger sample would have permitted this (Brown & Hudson, 2002). With a larger sample size, the relationships between test scores and questionnaire responses may have been explored using more complicated statistical techniques such as regression analysis, highlighting potential interactions between test taking behaviour and overall scores. Furthermore, collecting information about a larger group of participants would have supported exploration of biases in the test that were not immediately apparent from the test scores. Notwithstanding these limitations, the results of the research should be of interest to individuals responsible for developing L2 listening assessments in tertiary educational contexts. By evaluating an approach to replicate authentic speech using structured improvisation and developing tasks designed to elicit evidence of pragmatic competence using AI it is hoped this study provides some guidance to local language test developers and serves as an important caveat against utilising AI generated tests without systematic evaluation.

ORCID

https://orcid.org/0000-0003-3810-713X

Acknowledgements

Not applicable.

Funding

Not applicable.

Ethics Declarations

Competing Interests

No, there are no conflicting interests.

Rights and Permissions

Open Access

References

Aryadoust, V., Foo, S., & Ng, L. Y. (2022). What can gaze behaviors, neuroimaging data, and test scores tell us about test method effects and cognitive load in listening assessments? *Language Testing*, *39*(1), 56–89. https://doi.org/10.1177/02655322211026876

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

Brown, J. D. (2002). The Cronbach alpha reliability estimate. *JALT Testing & Evaluation SIG Newsletter*, *6*(1) 17-19.

Brown, J. D. (2006). Authentic communication: Whyzit importan' ta teach reduced forms?, *Authentic Communication: Proceedings of the 5th Annual JALT Pan-SIG Conference* (pp. 13 - 24). Tokai University College of Marine Science.

Brown, J. D. (2008). 11. Raters, Functions, Item Types and the Dependability of L2 Pragmatics Tests. In E. Alcón Soler & A. Martínez-Flor (Eds.), *Investigating Pragmatics in Foreign Language Learning, Teaching and Testing* (pp. 224-248). Multilingual Matters. https://doi.org/10.21832/9781847690869-013

Brown, J. D. (2014). *Mixed methods research for TESOL*. Edinburgh University Press. https://doi-org/10.1515/9780748698059-002

Brown, J. D. (2016). Assessment in ELT: Theoretical Options and Sound Pedagogical Choices. In W. Renandya & H. Widodo (Eds.), *English Language Teaching Today. English Language Education*, 5. Springer. https://doi.org/10.1007/978-3-319-38834-2_6

Brown, J. D., & Ahn, R. C. (2011). Variables that affect the dependability of L2 pragmatics tests, *Journal of Pragmatics*, *43*(1), 198-217. https://doi.org/10.1016/j.pragma.2010.07.026

Brown, J. D., & Crowther, D. (2022). *Shaping Learners' Pronunciation: Teaching the Connected Speech of North American English*. Taylor & Francis.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced Language Testing*. Cambridge University Press.

Brown, J. D., & Kondo-Brown, K. (Eds.). (2006). *Perspectives on teaching connected speech to second language speakers*. Natl Foreign Lg Resource Ctr.

Brown, J. D., & Trace, J. (2018). Connected-speech dictations for testing listening. In G. Ockey & E. Wagner (Eds.), *Assessing L2 Listening: Moving Towards Authenticity* (pp.46-63). John Benjamins.

Cardwell, R., LaFair, G., Naismith, B., & Settles, B. (2022). Duolingo English Test: Technical Manual. *Duolingo Research Report*. https://englishtest.duolingo.com/research

Carter, R., & McCarthy, M. (2006). *Cambridge grammar of English: A comprehensive guide: Spoken and written English grammar and usage*. Cambridge University Press.

Cheng, L., & Sultana, N. (2021). Washback Looking backward and forward. In G. Fulcher & L. Harding (Eds.), *The Routledge Handbook of Language Testing* (2nd ed., pp. 136-153). Routledge.

Clark, M. (2014). The use of semi-scripted speech in a listening placement test for university students. *Papers in Language Testing and Assessment*, *3*, 1–26.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press.

Council of Europe (2018). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors*. Council of Europe Publishing.

Dimova, S., Yan, X., & Ginther, A. (2020). *Local Language Testing Design, Implementation, and Development*. Routledge.

Field, J. (2011). Into the mind of the academic listener, *Journal of English for Academic Purposes*, *10*(2), 102-112. https://doi.org/10.1016/j.jeap.2011.04.002

Field, J. (2013). Cognitive validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining Listening* (pp.77-151). Cambridge University Press.

Green, A. (2020). *Exploring Language Assessment and Testing Language in Action*. Routledge.

Holzknecht, F., McCray, G., Eberharter, K., Kremmel, B., Zehentner, M., Spiby, R., & Dunlea, J. (2021). The effect of response order on candidate viewing behaviour and item difficulty in a multiple-choice listening test. *Language Testing*, *38*(1), 41–61. https://doi.org/10.1177/0265532220917316

Hudson, T., Detmer, E., & Brown, J. D. (1992). *A Framework for Testing Cross-Cultural Pragmatics*. University of Hawai'i Press.

Hudson, T., Detmer, E., & Brown, J. D. (1995). *Developing Prototypic Measures of Cross-Cultural Pragmatics (Technical Report # 7)*. University of Hawai'i at Manoa, Second Language Teaching and Curriculum Centre.

Isaacs, T., Hu, R., Trenkic, D., & Varga, J. (2023). Examining the predictive validity of the Duolingo English Test: Evidence from a major UK university. *Language Testing*, *40*(3), 748-770. https://doi.org/10.1177/02655322231158550

Ishihara, N., & Cohen, A. D. (2021). *Teaching and learning pragmatics*. Routledge.

Kaya, E., O'Grady, S., & Kalender, I. (2022). IRT-based classification analysis of an English language reading proficiency subtest. *Language Testing*, *39*(4), 541-566. https://doi.org/10.1177/02655322211068847

Khademi, A. (2023). Can ChatGPT and Bard generate aligned assessment items? A reliability analysis against human performance. *Journal of Applied Learning and Teaching*, *6*(1), 75-80. https://doi.org/37074/jalt.2023.6.1.28

Low, A. & Aryadoust, V. (2021). Investigating test-taking strategies in listening assessment: A comparative study of eye-tracking and self-report questionnaires. *International Journal of Listening*, *37*(2), 93-112. https://doi.org/10.1080/10904018.2021.1883433

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11.

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring, *Research Methods in Applied Linguistics*. *2*(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050

O'Grady, S. (2021). Adapting multiple-choice comprehension question formats in a test of second language listening comprehension. *Language Teaching Research*. Advance online publication. https://doi.org/10.1177/1362168820985367

O'Sullivan, B., Breakspear, T. & Bayliss, W. (2023). Validating an AI-driven scoring system: The Model Card approach. In K. Sadeghi & D. Douglas (Eds.), *Fundamental considerations in technology mediated language assessment* (pp.115-134). Routledge.

Roach, P. (1998). *English Phonetics and Phonology* (2nd ed.). Cambridge University Press.

Roever, C. (2011). Testing of second language pragmatics: Past and future. *Language Testing*, *28*(4), 463–481. https://doi.org/10.1177/0265532210394633

Settles, B., LaFlair, G., & Hagiwara, M. (2020). Machine Learning–Driven Language Assessment. *Transactions of the Association for Computational Linguistics*, *8*, 247–263. https://doi.org/10.1162/tacl_a_00310

Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E. (2018). Developing evaluative judgement: enabling students to make decisions about the quality of work. *Higher Education*, *76*, 467-481. https://doi.org/10.1007/s10734-017-0220-3

Taguchi, N. (2008). The effect of working memory, semantic access, and listening abilities on the comprehension of conversational implicatures in L2 English. *Pragmatics & Cognition*, *16*(3), 517-539. https://doi.org/10.1075/pc.16.3.05tag

Wagner, E. (2018). A comparison of listening performance on tests with scripted or authenticated spoken texts. In G. Ockey & E. Wagner (Eds.), *Assessing L2 listening moving toward authenticity* (pp. 29–44). John Benjamins.

Wagner, E., & Ockey, G. (2018). An overview of the use of authentic, real-world spoken texts on L2 listening tests. In G. Ockey & E. Wagner (Eds.), *Assessing L2 listening moving toward authenticity* (pp. 13–28). John Benjamins.

Youn, S., & Bi, N. (2019). Investigating test-takers' strategy use in task-based L2 pragmatic speaking assessment. *Intercultural Pragmatics*, *16*(2), 185-218. https://doi.org/10.1515/ip-2019-0009

Yunjiu, L., Wei, W., & Zheng, Y. (2022). Artificial intelligence-generated and human expert-designed vocabulary tests: A comparative study. *SAGE Open*, *12*(1). https://doi.org/10.1177/21582440221082130

Appendix A
Connected Speech Transcript

1. Yes nex' please

2. Er yes I' like t' return these er books erm I think they're due today I'm not sure

1. Ok that's grea' could I as' fe' your university ID number please

2. Yes its er nine one six er two zero five one eigh' two

1. Righ' I'll jus' log you into the system here er take the books thank you erm ri' that seems t' be fine yeah oh erm this las' book erm was due **last week**

2. Really

2. Yeah er it says tha' you checkt'it'out on the twenty seventh and it's a one week book so it should'erbeen returned'on the fourth

3. Ah cause I though' they were all due back on the same day er are you sure abou' that

3. Yes er the computer seems t' think so anyway erm I nee t' tell you there's a fine involved'of two pounds la' fee

4. Ok erm well I don' have money right now can I bring this afternoon

1. Yes of course th tha tha's fine there's no rush at'all

2. Ok grea sorry 'bout that I din' realise

1. That's ok

2. so wha's the normal erm loan perio'fe books

1. well erm it depens on the book erm ne relatively new books like this ten'to be one week only

ok I spose because lots o' people wanna borrow that book righ' ok tha' tha' makes sense ok