How well does the PTE-Academic test measure communication ability?

9 September 2024

Authors:

Ana Ulicheva

Sumita Ishaque

Rose Clesham

Contents

Abstract

In this study we measured interactional competence of 44 PTE Academic (PTE A research test) test-takers by means of three psycholinguistic tasks. Tasks consisted in describing abstract images, providing word definitions, and engaging in a real-life conversation with a peer. Students' performance on the three tasks was evaluated in terms of communicative success and correlated to PTE A scores. It was found that PTE A tasks generally tap into aspects of communicative ability. Our study provides initial evidence that PTE A measures important aspects on real-world communicative ability despite not including interactive items.

Introduction

Dialogue is one of the most common ways of using language. It involves using language in a collaborative manner to reach a common goal (Clark, 1996). Measuring how well someone engages in dialogue is important as it ensures that the language proficiency assessment reflects the real-world demands and uses of language. Interactive ability reflects the actual communicative demands test-takers will face, such as participating in meetings, engaging in academic discussions, or handling everyday interactions.

In broad terms, interactive ability is a skill that is more difficult to assess compared to reading, writing, and listening (de Jong, 2023; Fan & Yan, 2020). This is because speaking is context- and topic-dependent, multi-dimensional, and dynamic (Fulcher, 2015). Traditionally, assessments of speaking have relied on subjective (rated) or objective (e.g., number of errors) measures of lexical, grammatical, and fluency features (De Jong et al., 2012) characteristic of speech responses to prompts.

Measuring interactive ability in an authentic way is particularly challenging. It is well-known that a speaker's behaviour in a dialogical setting is heavily influenced by the type and the quality of input that their partner ('confederate') provides, even in situations when the confederate has been carefully trained and instructed to provide minimal or scripted input (Kuhlen & Brennan, 2013). In line with this, computerised high-stakes assessments, such as TOEFL or PTE A, involve no confederate. The downside is that, as some argue, such assessments only partially capture interactive communicative ability (Chalhoub-Deville, M., 2003; McNamara, 1996). Other tests like IELTS that involve a face-to-face interview and aim to test conversational skills (Taylor, 2001) involve minimal input from the confederate. In

this latter case, the interviewer's behaviour is heavily scripted and is designed to elicit more detailed extended responses that are again assessed based on their features (e.g., fluency and coherence, lexical resource), as opposed to direct measures of interactive ability (Brown, 2003; Fulcher, 2014). Another criticism of these types of tests is the predictability of topics and the use of rehearsed responses.

There have been other attempts to measure interactive or communicative ability in speaking assessment directly. Examples include a conversation with a virtual interlocutor or another candidate to navigate a role-play scenario, or complete a group task (e.g., Cambridge English Exams). Other approaches include task-based language assessments, where candidates are given tasks that simulate real-life situations, such as making a hotel reservation, giving directions, or discussing plans with a partner. Success is measured by how effectively candidates achieve the task's goal, considering both linguistic accuracy and communicative effectiveness. For example, if the task is to negotiate a meeting time, the assessment would look at whether the candidate can successfully agree on a time and place. To our knowledge, none of these approaches have yet been implemented in high-stakes contexts.

In psycholinguistic research, direct measurements of interactive ability (such as measures of turn-taking or common ground construction - Bovet et al., 2023; de Jong, 2023) are commonly utilised. For example, in referential communication tasks, one participant is asked to describe an object or picture, and the other is asked to identify it from a set of options. Success is measured by the accuracy of the identification and the efficiency of the description (Clarks & Wilkes-Gibbs, 1986). In collaborative tasks, instead, participants work together to complete a task, such as building a model or solving a puzzle. In this case, success is measured by the completion and quality of the final product (Barron, 2003).

In this paper, we adopted a psycholinguistic approach as described above, aiming to measure communication success via the use of more objective and direct measures. In particular, we investigate the extent to which typical language proficiency tasks such as those used in the computerised PTE A reflect test-takers' ability to communicate successfully.

Methods

We employed three tasks that are typically used to assess aspects of communication ability.

The first task, that we refer to as *tangrams*, is a variation on the canonical referential communication task that involves the presentation of abstract figures to

participants (Clark & Wilkes-Gibbs, 1986; Lev-Ari & Sebanz, 2020). A participant ('speaker') has to refer to tangrams to help another person ('listener') arrange them in a correct order. The use of tangrams is particularly relevant in studying common ground construction and use between two speakers, - as dialogue partners are usually unfamiliar with the tangrams' shapes, i.e., they have no common ground regarding how to refer to them (Bovet et al., 2023). However, one issue with using this task is that the speaker's behaviour is affected by the type and quality of input that the listener provides (Kuhlen & Brennan, 2013). In designing this task, we drew inspiration from Lev-Ari & Sebanz (2020) who separated the speaker and listener turns in time. In other words, we collected speakers' descriptions at timepoint 1, and presented them to a listener at timepoint 2 to determine communication success (i.e., whether the descriptions were good enough for the listener to reconstruct the correct order).

In the second task (*the definitions task*), speakers provide definitions for abstract and concrete terms. Word definition tasks can be used to assess vocabulary depth and the ability to provide accurate definitions. Performance on this task was found to be correlated with language proficiency (Rosqvist et al., 2022). The ability to produce informative, relevant definitions depends on a person's semantic knowledge, their linguistic ability, metalinguistic skills (Bialystok & Ryan, 1985; Astell & Harley, 2002; Marinellie & Johnson, 2004), and pragmatic ability (Gutierrez-Clellen & DeCurtis, 1999). Our version of this task asked a group of listeners to identify the referent being described by the speakers.

The third task involved an actual *conversation* between two study participants whose level of proficiency was similar.

Prior to completing the three tasks, speakers took a global test of English language proficiency (research version of PTE A, see below). To address our research questions, we calculated correlations between psycholinguistic measures derived from the three tasks and the scores obtained on a set of speaking tasks from PTE A.

Participants

Participants were 44 current undergraduate or postgraduate university students in a UK institution. We recruited students with any language background. Half of the participants were born in China, and the rest came from Europe, Africa, Middle East, and Asia. Consequently, 59% indicated that their first language was Chinese, 32% English, and the remaining 9% spoke Arabic, Japanese or Russian as their first language.

Participants were, on average, 23 years old (ranging between 19 and 37); 34 identified as females and 10 as males.

Regarding students' institutions, one-third came from University College London, 35% from Southampton University, and the rest came from London School of Economics, Royal Holloway University of London and other institutions in the country.

Materials

PTE Academic research test. The PTE Academic exam evaluates candidates' English proficiency within an academic context to determine their preparedness for studying in English-speaking environments. Conducted entirely on computers at Pearson VUE test centres, the test combines AI automated scoring and human expertise. Each test taker receives a comprehensive score report that includes an overall score, and individual scores for listening, reading, speaking, and writing. PTE A tasks are integrated task types, meaning that they require the combined use of multiple communicative skills, e.g., listening and speaking, or listening and writing. A research version of the PTE A test was administered. Using research versions allows us to explore particular areas of interest, such as field testing new or amended item types and rubrics. In this case, the research version was largely unchanged from the operational test in its composition, scoring, or administration. The differences concerned minor modifications to selected items (e.g., to have more word count), the addition of new items (summarise group discussion, respond to a situation), and changes to some item's scoring rubrics.

Remote tasks. Remote psycholinguistics tasks were created and hosted on the Gorilla platform (https://gorilla.sc/; Anwyl-Irvine et al., 2019) – a service widely used among academics in cognitive psychology and life sciences to run behavioural experiments. Participants received a link and accessed the tasks via a web browser on their PC or laptop. No other devices, such as phones or tablets, were allowed. Aside from tasks that we report on in this paper, participants completed a range of other language tasks. The total duration of the whole suite of tasks was between 50 and 70 minutes. Short breaks in-between tasks were allowed. A quiet environment was recommended.

Questionnaire. A student questionnaire was administered before starting the tasks. We collected data on participants' demographic and language background. Questions were adapted from the Language Experience and Proficiency Questionnaire; Kaushanskaya et al., 2020).

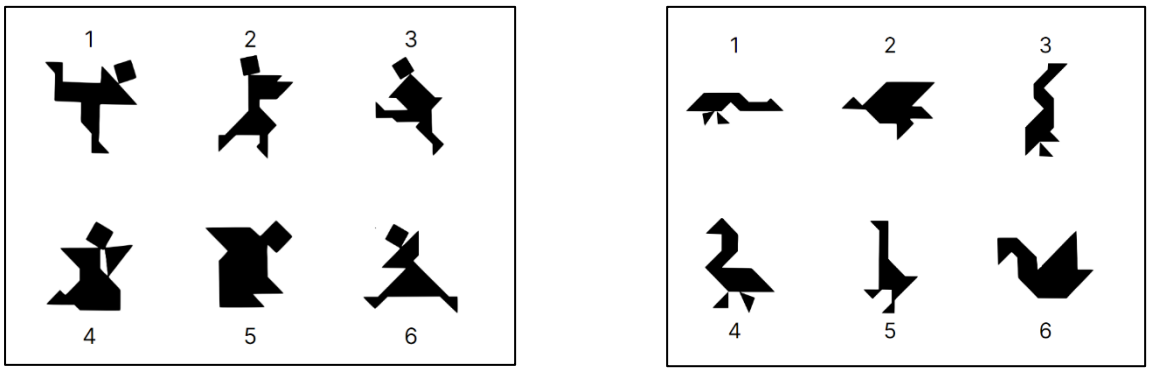
Tangrams task

This task involved a two-phased data collection approach. In Phase I, participants had to describe the order of pictures, and their descriptions were recorded. In

Phase II, these descriptions were played to a group of raters who attempted to reconstruct the correct order of pictures.

Phase I. We selected two sets of 6 pictures from Fasquel et al. (2023), one set resembling a dancing human, the other set resembling geese. Pictures within sets were highly similar to each other so that participants would need to provide a lot of detail when describing them.

Figure 1. Tangram figures sets: human-like on the left and goose-like on the right.



Participants were told that their descriptions would be presented to another person, and they would be rated on how accurately that person performs the order-reconstruction task (see the full set of instructions in Appendix). The task consisted of two blocks, each including four trials. The order of blocks was randomised. On each trial, participants saw 6 pictures from one set (e.g., geese-like pictures), arranged over two rows and numbered from 1 to 6. The order of pictures was random. Their descriptions were audio-recorded. Time for responding was unrestricted. The duration of this task was around 15 minutes.

Phase II. In a separate experiment, 10 raters listened to the recordings and arranged pictures in front of them in the order that they felt the speaker was trying to convey. Arranging was done via the drag-and-drop mechanism. Once dragged into place, pictures could not be moved. Participants also had an opportunity to report empty or incomplete recordings. Participants were recruited via Prolific (prolific.com). Session duration was 20-30 minutes on average (participants also had to complete Phase II of the definitions task within the same session). Participants were paid at £12.75 per hour. 420 native speakers of English from the UK, USA, Ireland, Australia, and Canada were recruited, such that 10 participants evaluated each person's responses from Phase I. Approval rate of participants on the platform was set above 90% to ensure data quality. Further, manual review of submissions was undertaken following data collection (participants who failed accuracy checks were rejected). Each recording was auto-played at least once, but replays were permitted.

Definitions task

This task involved a two-phased data collection approach. In Phase I, participants generated definitions for a set of words. In Phase II, these definitions were played to a group of raters who evaluated them in terms of their communicative adequacy.

Phase I. Sixteen words were selected for this task. This number was chosen as Rosqvist et al. (2022) reported that high internal validity can be achieved for a 10-item task of this nature. Further, we controlled for a number of psycholinguistic variables that are known to affect the difficulty of word definition (frequency, concreteness, semantic richness). In terms of frequency, half of the selected words were high-frequency words (above 9.36, based on logarithm-transformed words-per-million HAL frequencies, Balota et al., 2007), the other half were low-frequency words (below 8.06). The second variable was concreteness (Kim et al., 2015; McGregor et al., 2012). Concrete nouns were defined as having values over 5.15, abstract nouns were defined as having values below 3.81. Finally, we counted the number of core semantic features for these words based on the dataset provided by Buchanan et al. (2017). The number of core semantic features is another dimension of semantic richness and affects the quality of definitions (Astell & Harley, 2002; Buchanan et al., 2017).

Table 1. Target words used in the definitions task with their psycholinguistic characteristics.

Frequency Code	Concreteness	Feature set	Word	Frequency (HAL)	Concreteness Value	Number of features
HF	Concrete	Large	castle	9.36	6.5	14
HF	Concrete	Large	beer	10.11	5.83	7
HF	Concrete	Small	dragon	10.29	5.49	4
HF	Concrete	Small	music	11.81	5.15	5
HF	Abstract	Large	sale	11.47	3.6	10
HF	Abstract	Large	language	11.49	4	8
HF	Abstract	Small	crime	10.42	3.81	4
HF	Abstract	Small	memory	11.65	1.78	4
LF	Concrete	Large	napkin	5.76	5.29	16
LF	Concrete	Large	skyscraper	5.24	6.14	11
LF	Concrete	Small	yolk	5.78	5.89	4
LF	Concrete	Small	deodorant	5.18	6.19	5
LF	Abstract	Large	yacht	6.91	3.35	10
LF	Abstract	Large	bravery	6.12	1.93	7
LF	Abstract	Small	autumn	8.06	3.5	3
LF	Abstract	Small	intuition	7.51	2.3	4

Words were presented in the centre of the screen, one at a time. The maximum of 30 seconds was allowed to define each word, but participants could move forward if

they finished earlier. A countdown clock was displayed on each trial. Instructions were adapted from Kim et al. (2015; see Appendix). Participant's vocal descriptions were audio-captured as individual recordings. One practice word was presented first ("apple"), followed by 16 experimental words presented in a random order for every individual. The duration of this task was around 10 minutes.

Phase II. In a separate experiment, 10 raters rated communicative adequacy of each definition collected in Phase I. The procedure was similar to that used for the Phase II of the tangrams task. Raters listened to each recording and were asked to identify the referent of the definition by typing their guess(es) into the designated fields. After that, they were shown the correct referent and were asked to score the definition based on the following rubrics: 0 repetition without explanation or omission, no response, incorrect definition; 1 bears some meaningful relationship to the target but does not define it; 2 several targets could be possible; 3 one or two targets seem most likely; 4 this directed me to the target (accurate and well-elaborated amount of information). We provided an example of possible definitions of "carrot": "a thing with four legs" (rated 0), "a vegetable" (rated 1), "a root vegetable" (rated 2), "a long root vegetable" (rated 3), and "a long orange-coloured root vegetable" (rated 4; Astell & Harley, 2002; Rosqvist et al., 2022).

Phase II participants listened to all definitions generated by one person one at a time, presented in a random order. Each definition was auto-played at least once, but replays were permitted. Four boxes for typing their guesses were provided on the same screen. Two additional checkboxes were present on the same screen: one for reporting recordings that mention target words and the other for reporting difficulties with guessing. Next, participants moved on to the next screen where they again had an opportunity to listen to the same definition as many times as they wished. They then had to choose one option from a drop-down menu to indicate the quality of the definition (see above). Note that the recordings that participants evaluated in Phase II of the tangrams task belonged to a different person that the recordings they evaluated in Phase II of the definitions task.

Conversations task

We selected three topics for conversation: film preferences (Rimé, 1982), close-call incidents (Bavelas, Coates, & Johnson, 2000), and the resolution of an ethical dilemma (Healey, Purver, King, Ginzburg, & Mills, 2003). Topics were displayed for the participants alongside detailed prompts to guide the conversations. The reason for selecting three topics with different context, register, and communicative demands was primarily to provide participants with ample opportunities to showcase their communicative skills, which might lie in different domains, for example, one person's strength may lie in informal communication, while another person might be skilled at argumentative academic-like discussions. We did not

subsequently differentiate between the topics in our analyses; all measures were averages across the whole conversation.

For film preferences, they were prompted to “discuss this topic, explain to each other their opinions on movies, and express what they like to find in the cinema”. For the close-call accident topic, the following prompt was presented: “What I'd like both of you to do is tell something about a close call or near-miss incident. A close call is something that happened where someone was almost hurt, or something bad almost happened, but in the end everything turned out okay. Make sure that you tell something you're comfortable telling. And if you can't think of something that happened to you, then you can tell about a close call that happened to a friend. Just to give you some ideas, other people have told stories about skiing accidents, horseback-riding accidents, and nearly losing a paper on the computer. I would like you to tell your story in as much detail as you can. So don't just describe it in a couple of sentences.” Finally, the ethical dilemma was prompted using the following story: “A balloon is losing altitude and about to crash. The only way for any of three passengers to survive is for one of them to jump to a certain death. The three passengers are Dr. Nick Riviera, a cancer scientist, Mrs. Susie Derkins, a pregnant primary school teacher, and Mr. Tom Derkins, the balloon pilot and Susie’s husband Decide who should jump!”

We divided our sample into dyads based on the students’ level of proficiency (as indicated by PTE A) and their age. Some dyads were same-gender, others were mixed-gender, depending on availability. Prior to the experiment, participants were strangers to each other (except for one pair; see Hadley et al., 2021, for a similar approach). We ran the sessions individually via Teams, with at least one experimenter present. The experimenter introduced the participants to each other by their first name, shared Power Point slides describing the instructions for the task (see Appendix) and introducing the topics one-by-one. If the duration of each dialogue exceeded five minutes, the experimenter moved on to the following slide indicating that the time was up. The whole session was recorded. The advantage of online administration was that it allowed for time-stamped automatic transcription, which reduced the workload on the team. The duration of this task was kept under 20 minutes.

Procedure

After students expressed their interest, all received instructions via email on how to prepare for and sit the PTE A. The test was administered in one of the UK centres in person. The cost of PTE A was waived. Participants were compensated for their travel to and from their closest test centre. They were offered additional remuneration for completing all requirements for the study at an hourly rate exceeding minimum wage in the UK. Data collection took place between the 8th of

January and the 8th of March 2024. Participants were also offered a free PTE A voucher. There were not given any feedback on their performance in the PTE A test.

After participants sat PTE A, they received instructions on how to complete remote tasks. They were urged to complete the tasks as soon as possible. All participants received detailed information on Pearson's privacy policy and data handing and use, and they were asked to give consent to participate in the study.

Analysis

Tangrams measure. Participants' behaviour on this task was evaluated by raters in Phase II. In particular, each Phase I participant's response was evaluated in terms of the number of correct placements made by listeners (ranging from 0, i.e., no correct placements, to 6, i.e., all pictures were positioned correctly). Participant-level measures were obtained by averaging these two metrics across the four trials. We refer to this measure as 'the number of correct placements in the tangrams task'.

Definitions measures. The definitions were evaluated in Phase II. We calculated average rating that Phase II raters assigned to each definition (from 0 to 4), proportion of participants that guessed the target word, and semantic similarity of guesses to the target word. For the latter two measurements, we identified and corrected erroneous spellings of guesses first. This was done by applying the *hunspell()* function from the 'hunspell' package (Ooms, 2014) in R to lowercase spellings and adopting the first most likely spelling suggestion for cases that were identified as errors. Semantic similarity of guesses to the target word was calculated using the English LSA semantic space with 300 dimensions that was created from a 2.8 billion word corpus (combining the British National Corpus (BNC), the ukWaC corpus and a 2009 Wikipedia dump; Günther et al., 2015). As each Phase I participant had to define 16 words, the resulting participant-level measures were averages across these 16 definitions.

Conversations measures. The transcribed recording was analysed to yield two measures for each participating speaker. Start and end time for each speaking turn was extracted using automatically generated time stamps. Number of words per turn was calculated. Experimenter's turns were removed from the analysis. The first measure, pace, represented the total number of words produced by the speaker across all their turns adjusted for their total speaking time. The second measure, number of turns, was the total number of speaking turns taken by the speaker adjusted for the recording duration.

PTE A measures. We considered the following measures from the PTE Academic: overall score, skill subscores (listening, speaking, reading, writing), as well as raw scores on all individual items (machine-generated). Our main analysis involved computing Pearson's correlations between each psycholinguistic measure and (1)

overall PTE Academic scores, as well as (2) individual item scores. In the section below we report these correlations separately for each psycholinguistic task.

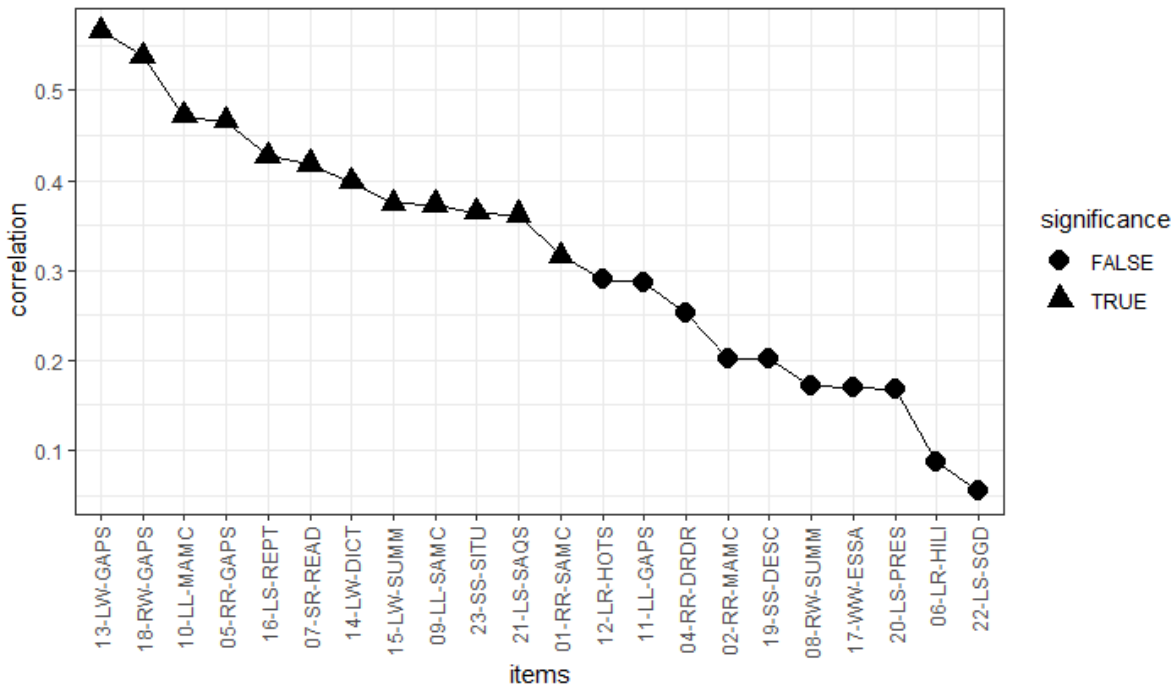
Results

Tangrams

Table 2. Correlation between overall and section-level PTE scores and the number of correct placements measure from the tangrams task.

	correlation	significance
Overall.Score	0.54	0.0002
Listening.Score	0.52	0.0005
Reading.Score	0.56	0.0001
Speaking.Score	0.41	0.0064
Writing.Score	0.54	0.0002

Figure 2. Correlation between item-specific scores and the number of correct placements measure from the tangrams task.



Definitions

Figure 3. Correlation between overall and section-level PTE scores and psycholinguistic measures derived from the definitions task: avg.rating is the average rating assigned to definitions, on average, in Phase II; target.guesses is the proportion of participants who listed the target word among their guesses; semantic.similarity is the average semantic similarity of all guesses to the target word.

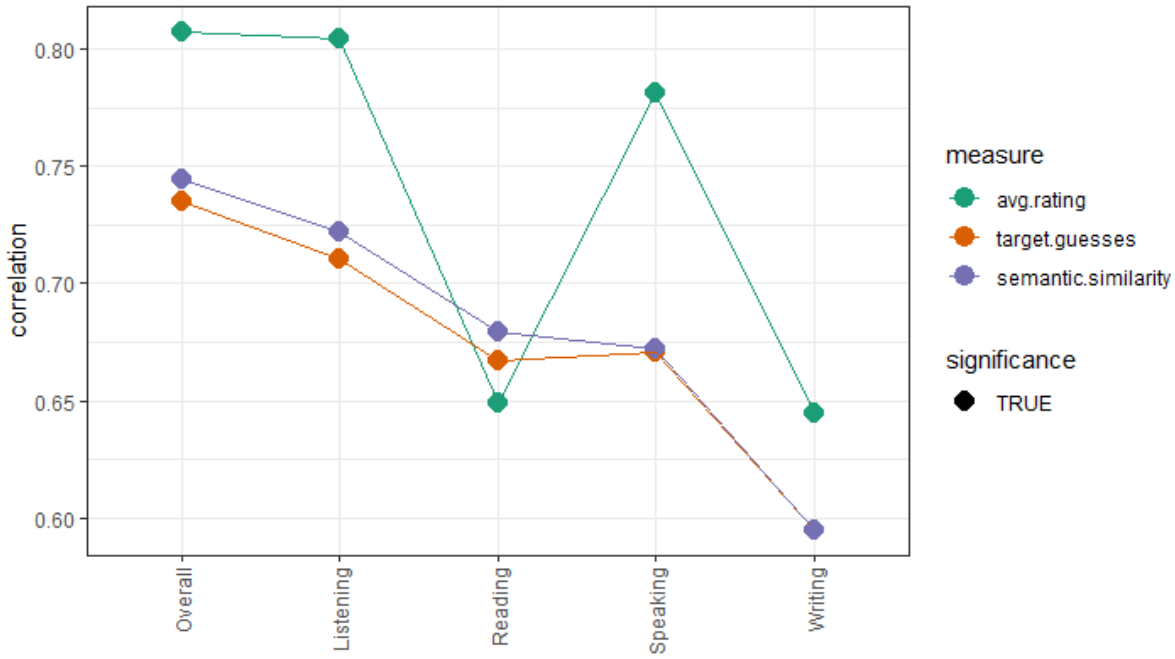
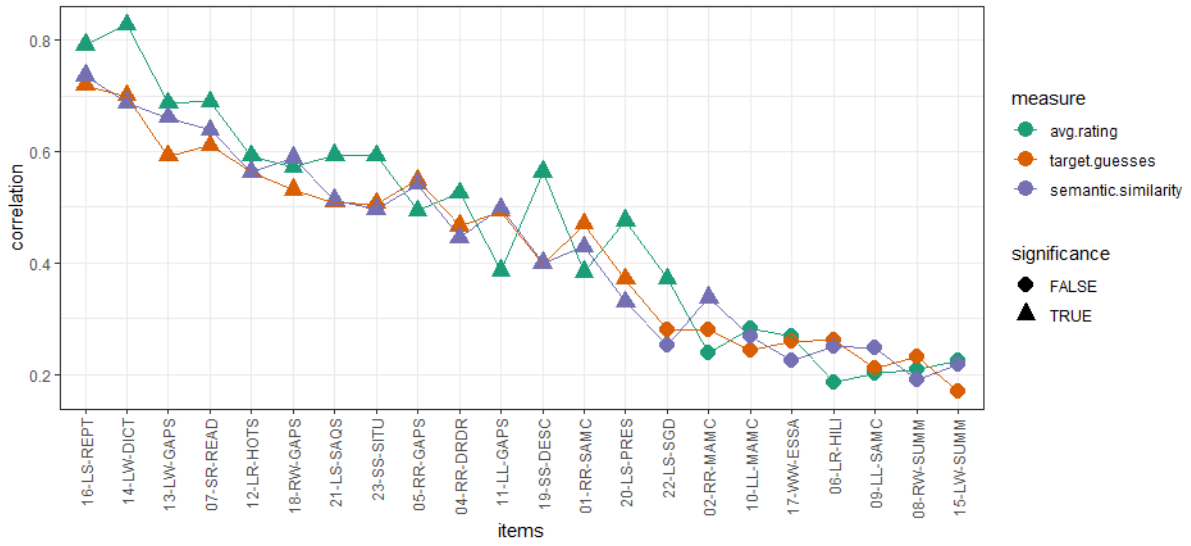


Figure 3. Correlation between item-specific scores and psycholinguistic measures derived from the definitions task: avg.rating is the average rating assigned to definitions, on average, in Phase II; target.guesses is the proportion of participants who listed the target word among their guesses; semantic.similarity is the average semantic similarity of all guesses to the target word.



Conversations

Figure 4. Correlation between overall and section-level PTE scores and psycholinguistic measures derived from the conversation task: pace is the total number of words produced by the speaker adjusted for their total speaking time; n.turns is the number of speaking turns taken by the speaker adjusted for the recording duration.

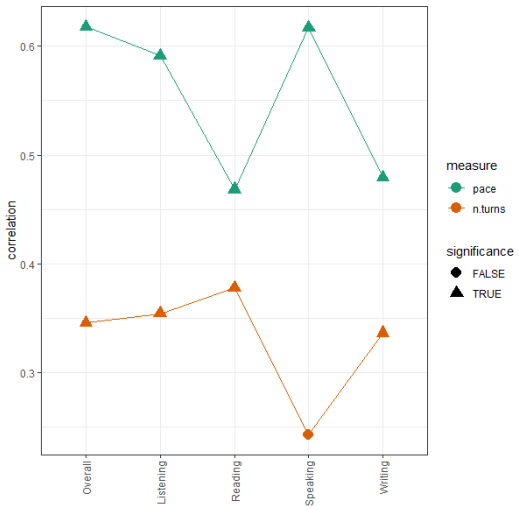
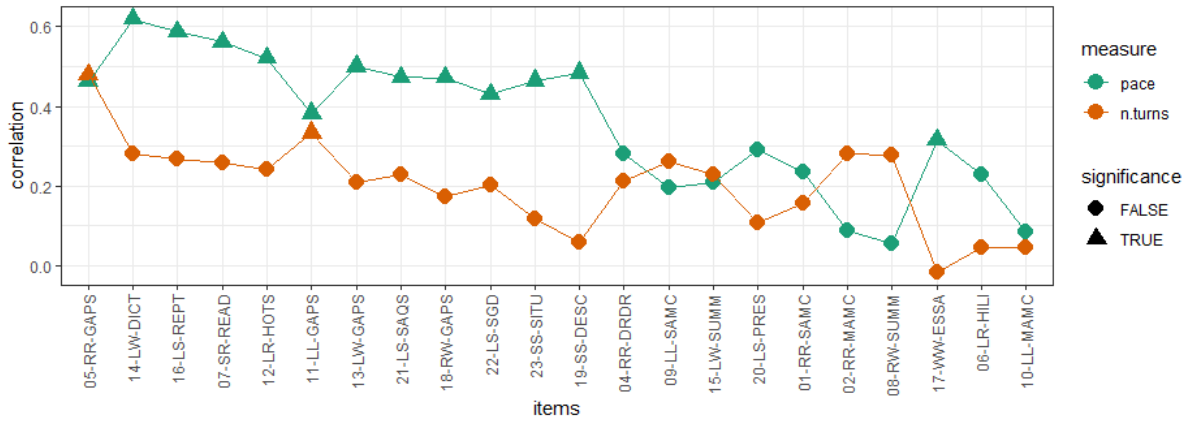


Figure 4. Correlation between item-specific scores and psycholinguistic measures derived from the conversation task: pace is the total number of words produced by the speaker adjusted for their total speaking time; n.turns is the number of speaking turns taken by the speaker adjusted for the recording duration.



Discussion

'Read Aloud' and 'Repeat Sentence'

The first finding that emerged from these data was that items 'Read Aloud' 07-SR-READ and 'Repeat Sentence' 16-LS-REPT were highly correlated with our communication measures, i.e., tangrams, definitions, and conversations.

'Read Aloud' is designed to measure pronunciation, fluency, and reading comprehension. The item description states: "The purpose of this task is to measure 'manner of speaking' skills that are foundational to oral communication." Indeed, our findings point to a clear link between this item and communicational competence.

Similarly, 'Repeat Sentence' is designed to measure speech perception and speech production processes (pronunciation, fluency), as well as aspects of speaking proficiency. The item description states: "Test takers must understand spoken utterances as well as recognize and process linguistic units in order to repeat a sentence, just as they need to do in real-time conversations or other listening and speaking situations in the real world". Previous research has shown a correlation between repeat sentence tasks and oral narrative tasks (Sarandi, 2015), lexical diversity in oral interviews and speech rate (Tracy-Ventura, McManus, Norris, & Ortega, 2013), and pronunciation, fluency and content ratings in story retell and respond to situation tasks (Van Moere, Xu, & Klungtvedt, 2012, unpublished). Our study is the first to provide direct evidence that performance on the PTE A 'Repeat Sentence' task is relevant to skills underlying communication ability.

'Fill in the Gaps' ,'Write to Dictation', and 'Select Missing Word'

These items were also sensitive to aspects of communicational measurements. Unlike 'Read Aloud' and 'Repeat Sentence', their description does not mention communication. Clearly, based on our findings, these items also tap into aspects of

comprehension and vocabulary knowledge that are important in successful oral communication. Of particular interest are correlations between 'Fill in the Gaps'/' Select Missing Word' and the psycholinguistic measures from the conversations task. Arguably, this is the task that has the highest face validity when it comes to interactional competence. 'Fill in the Gaps' and 'Select Missing Word' were significantly related to the number of turns that participants exchanged during the live conversation. This suggests that semantic skills tested in the 'Fill in the Gaps' (reading) item and oral comprehension skills tested in the 'Select Missing Word' item predict to some extent how successfully a test taker would engage in a real-life conversation on a given topic.

'Respond to a Situation' and 'Summarise Group Discussion'

These new items performed well with regards to some psycholinguistic measures (especially those from the definitions and conversations task). Summarise Group Discussion item type is intended to assess listening comprehension of multiple perspectives in dialogue, ability to synthesize information, express ideas precisely, organise ideas logically, and communicate this in spoken English. 'Respond to a Situation' is evaluated based on the success of the response in achieving the primary goal of the communication (e.g., apology, request etc) while taking into account the context of the situation given in the prompt. It is reassuring to see these two items that are designed to assess aspects of communicative ability, indeed are well aligned with parallel measures of this ability.

'Retell Lecture'

Finally, 'Retell Lecture' was less sensitive than other speaking items to some aspects of communicational competence – although this item is meant to test pronunciation and fluency, as well as mimic real-life **domain-relevant contexts, such as participating in a follow-up Q&A or discussion session. Findings from this study suggest we should carry out additional research to investigate the construct definition and scoring approach for this task, e.g., whether this item taps different dimensions that have not been included in the definition.**

Broader implications of this work include several points. Firstly, PTE A tasks capture aspects of communicative competence despite being computerised and monologic in nature. While it is generally believed that extended responses are superior in measuring this construct (Chalhoub-Deville, M., 2003; McNamara, 1996), our study demonstrates that lower-effort, simpler tasks are remarkably good indicators as well (see Van Moere, 2012; de Jong, 2023, Davis & Norris, 2021, for a similar argument).

Secondly, PTE A tasks can be generalised to real-world communicative situations. Our tasks involved a separate group of English raters listening and evaluating test-taker responses. Candidates had to make themselves understood by speaking in a comprehensible manner and providing enough detail for others to infer what they were referring to. Such measures are rarely used for assessing communicative

competence for they are extremely resource-demanding. Our approach lends validity to the types of items used in PTE Academic – as we could demonstrate that those who score well on these items are also able to achieve a variety of communicative goals.

Lastly, we found that measurement of communicative ability is distributed across the whole test, and listening, reading and writing tasks pick up relevant aspects of communicational competence alongside speaking tasks (see also Ricketts, 2011; Gottardo & Mueller, 2009; Lynch, 1997).

References

Astell, A. J., & Harley, T. A. (2002). Accessing semantic knowledge in dementia: evidence from a word definition task. *Brain and Language*, *82*(3), 312-326.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., et al. (2007). The English Lexicon Project. *Behavior Research Methods, 39*, 445-459.

Barron, B. (2003). When smart groups fail. *The journal of the learning sciences*, *12*(3), 307-359.

Bavelas, J. B., Coates, L., & Johnson, T. (2000). Listeners as co-narrators. *Journal of personality and social psychology*, *79*(6), 941.

Bialystok, E., & Ryan, E. B. (1985). Toward a definition of metalinguistic skill. Merrill-Palmer Quarterly (1982-), 229-251.

Bovet, V., Knutsen, D. & Fossard, M. Direct and indirect linguistic measures of common ground in dialogue studies involving a matching task: A systematic review. Psychon Bull Rev (2023). https://doi.org/10.3758/s13423-023-02359-2

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language testing*, *20*(1), 1-25.

Buchanan, E.M., Valentine, K.D. & Maxwell, N.P. English semantic feature production norms: An extended database of 4436 concepts. *Behav Res* 51, 1849–1863 (2019). https://doi.org/10.3758/s13428-019-01243-z
https://doomlab.shinyapps.io/single_words/

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language testing*, *20*(4), 369-383.

Clark, H. H. (1996). Using language. Cambridge university press.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. Cognition, 22(1), 1-39.

Council of Europe (2020). Common European Framework of Reference for Languages: Learning, Teaching, Assessment—Companion Volume. Council of

Europe Publishing. Qualitative aspects of spoken language use – Table 3. Retrieved from https://www.coe.int/en/web/common-european-framework-reference-languages/table-3-cefr-3.3-common-reference-levels-qualitative-aspects-of-spoken-language-use

Davis, L., & Norris, J. (2021). Developing an innovative elicited imitation task for efficient English proficiency assessment. *ETS Research Report Series*, *2021*(1), 1-30.

De Jong, N. (2023). Updating the construct of speaking proficiency. Paper presented at the 44th Language Testing Research Colloquium.

de Jong, N. H. (2023). Assessing second language speaking proficiency. *Annual Review of Linguistics*, *9*, 541-560.

De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in second language acquisition*, *34*(1), 5-34.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological review*, *104*(4), 801.

Fan, J., & Yan, X. (2020). Assessing speaking proficiency: a narrative review of speaking assessment research within the argument-based validation framework. *Frontiers in psychology*, *11*, 513393.

Fasquel, A., Brunellière, A. & Knutsen, D. A modified procedure for naming 332 pictures and collecting norms: Using tangram pictures in psycholinguistic studies. Behav Res 55, 2297–2319 (2023). https://doi.org/10.3758/s13428-022-01871-y

Fulcher, G. (2014). *Testing second language speaking*. Routledge.

Fulcher, G. (2015). Assessing second language speaking. *Language Teaching*, *48*(2), 198-216.

Gottardo, A., & Mueller, J. (2009). Are first-and second-language factors related in predicting second-language reading comprehension? A study of Spanish-speaking children acquiring English as a second language from first to second grade. *Journal of Educational Psychology*, *101*(2), 330.

Gutierrez-Cleflen, V. F., & DeCurtis, L. (1999). Word definition skills in Spanish-speaking children with language impairment. Communication Disorders Quarterly, 21(1), 23-31.

Hadley, L. V., Whitmer, W. M., Brimijoin, W. O., & Naylor, G. (2021). Conversation in small groups: Speaking and listening strategies depend on the complexities of the environment and group. *Psychonomic Bulletin & Review*, *28*, 632-640.

Healey, P. G., Purver, M., King, J., Ginzburg, J., & Mills, G. J. (2003). Experimenting with clarification in dialogue. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 25, No. 25).

Kim, S. R., Kim, S., Baek, M. J., & Kim, H. (2015). Abstract word definition in patients with amnestic mild cognitive impairment. Behavioural neurology, 2015.

Kuhlen, A. K., & Brennan, S. E. (2013). Language in dialogue: When confederates might be hazardous to your data. Psychonomic bulletin & review, 20, 54-72.

Lev-Ari, S., & Sebanz, N. (2020). Interacting with multiple partners improves communication skills. Cognitive Science, 44(4), e12836.

Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and brain sciences*, *22*(1), 1-38.

Lynch, T. (1997). Life in the slow lane: Observations of a limited L2 listener. *System*, *25*(3), 385-398.

Marinellie, S. A., & Johnson, C. J. (2004). Nouns and verbs: A comparison of definitional style. Journal of psycholinguistic research, 33, 217-235.

McGregor, K. K., Berns, A. J., Owen, A. J., Michels, S. A., Duff, D., Bahnsen, A. J., & Lloyd, M. (2012). Associations between syntax and the lexicon among children with or without ASD and language impairment. *Journal of autism and developmental disorders*, *42*, 35-47.

McNamara, T. F. (1996). Measuring second language performance. London: Longman

Ricketts, J. (2011). Research review: Reading comprehension in developmental disorders of language and communication. *Journal of Child Psychology and Psychiatry*, *52*(11), 1111-1123.

Rimé, B. (1982). The elimination of visible behaviour from social interactions: Effects on verbal, nonverbal and interpersonal variables. *European journal of social psychology*, *12*(2), 113-129.

Rosqvist, I., Andersson, K., Sandgren, O., Lyberg-Åhlander, V., Hansson, K., & SahlÉn, B. (2022). Word definition skills in elementary school children–The contribution of bilingualism, cognitive factors, and social factors. International Journal of Speech-Language Pathology, 24(6), 596-606.

Schmitt, N. (2010). Researching vocabulary: A vocabulary research manual. Springer.

TAYLOR, L. (2001). THE PAIRED SPEAKING TEST FORMAT: RECENT STUDIES. RESEARCH NOTES, 6(-), 15-16. SID. https://sid.ir/paper/615390/en

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing*, *29*(3), 325-344.

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. Applied psycholinguistics, 22(2), 217-234.

Appendix

Instructions for each task. Paragraphing not preserved here.

Definitions instructions. In this task, you will be presented with a single written word. Please explain what this word means without mentioning the word itself. Give as much information about the word as possible. The aim is that the word is recognisable from the definition alone.

There is a small prize (£10) for the test-taker who provides the most useful definitions. This will be judged by an independent person listening to definitions and trying to guess the word: the individual with the highest number of correctly guessed words wins.

The focus on this experiment is on how the meaning of words can relayed, so please focus on how you could define the word and avoid very obvious measures such as spelling the word, or giving clues to how the word might sound. For example, if the word on the screen is "hat", avoid saying "it is a three-letter word beginning with H" and/or "it rhymes with cat". Such definitions will not score any points.

There are 17 words in total. You have up to 30 seconds to describe each word. You don't have to fill all 30 seconds if you feel that your definition is good enough.

Tangrams instructions. In this task you are provided different sets of pictures. In each set, you will see six abstract pictures. These are arranged in a specific order. You need to describe each picture in the order that you see these. The idea is that a separate person will be given the same abstract pictures but in a different order. From just listening to your descriptions in your recording, they will attempt to reconstruct the correct order (i.e. the order seen by you). Make sure that the description includes enough detail for this person to accomplish this task! You will be given four separate sets. Please treat these as completely discrete. Why? When describing one set, you may see an abstract picture that you have already seen on another set. However, your audience will not necessarily see the sets in the same order as you, so please avoid referring directly to an earlier set. Just stick with describing the abstract picture. The audience will not always be the same. We will define which audience you are speaking to, for each set.

Conversations instructions. You'll see one topic per slide. There are three topics in total. For each topic, talk with each other for up to five minutes. Prompts are provided but these are not exhaustive. Feel free to talk about anything on a given topic. Act as naturally as possible, just as if talking to a friend. Do engage with each other as much as you can.