Final Report

Project Title	Validating the Speaking and Writing Sections of PTE Academic: Performance and Perceptions of International Students in Freshman Writing Courses
Reporting Period	August 2024 – November 2025
Principal Investigator	Hyung-Jo Yoon, California State University, Northridge
Co-PI	Daniel R. Isbell, University of Hawai‘i at Mānoa

Introduction

Standardized assessments of English language proficiency play a central role in international student admissions and placement decisions at English-medium universities. Among the available assessments, the Pearson Test of English (PTE) Academic has gained visibility due to its technological accessibility, relatively short administration time, and fully automated scoring system (Pearson, 2021). While these features contribute to test practicality, they also raise questions regarding the nature and validity of the evidence produced, particularly for productive language skills. In contrast to other highly computerized tests with more limited academic content (e.g., such as the Duolingo English Test; see Wagner, 2020), PTE Academic includes a series of open-ended speaking and writing tasks designed to elicit contextualized academic discourse. These tasks aim to measure linguistic and discourse abilities relevant to university settings, such as describing information from visual and oral sources, synthesizing ideas, and producing extended written arguments.

A central question in test validation concerns the degree to which test tasks represent the linguistic and cognitive demands of the target language use (TLU) domain (Bachman & Palmer,

2010; Kane, 2013). Test developers and test users expect that performance on PTE Academic and similar tests used for admissions decisions will generalize to students' academic speaking and writing in university contexts, yet empirical evidence demonstrating clear alignment remains limited. Existing validation research often focuses on relationships between proficiency test scores and course grades or on comparisons between standardized tests (e.g., Biber et al., 2017; Riazi, 2013; Llosa & Malone, 2019). However, performances on test tasks and authentic course assignments may not always correspond, and textual quality alone provides an incomplete picture of language ability (Crossley & McNamara, 2014; Deane, 2013). A broader argument-based approach to validity requires evidence not only that the test predicts academic outcomes (e.g., GPA, grades on course assignments), but also that test responses share linguistic and cognitive characteristics with authentic academic performances (Chapelle, 1999; Kane, 2013).

Productive language tasks are often evaluated according to multiple dimensions, including syntactic complexity, lexical complexity, accuracy, and fluency (Bulté & Housen, 2012). These linguistic measures are thought to reflect task difficulty and are indicative of cognitive processing demands and underlying linguistic competence (Robinson, 2001) and have been used to examine alignment between standardized test tasks and academic assignment genres (e.g., Riazi, 2016). Perceptions of task difficulty and relevance also provide insight into the cognitive demands of test tasks, because they reflect the degree to which tasks elicit familiar or unfamiliar performance processes (Brooks & Swain, 2014; Cumming et al., 2006).

Motivated by these considerations, the present study examines the performance, linguistic characteristics, and perceptions related to PTE Academic Speaking and Writing tasks among international undergraduate students enrolled in a required first-year writing course. This study adopts three guiding research questions:

1. To what extent do scores on PTE Academic speaking and writing tasks predict the quality of language performance on authentic speaking and writing tasks?
2. To what extent do the linguistic characteristics of PTE Academic speaking and writing task responses correspond to those of authentic academic speaking and writing tasks?
3. How do international students enrolled in a required writing course perceive productive PTE Academic tasks?

By combining score-based comparisons, linguistic analyses, and student perception data, this study contributes important evidence to evaluate claims about the meaning and relevance of PTE Academic scores. The findings provide insight into alignment between standardized test tasks and university academic demands and inform interpretations of their productive skill scores. Nevertheless, it should be noted that the data analyzed in the present study were collected prior to the implementation of broader updates to PTE Academic. Since August 2025, PTE Academic has operationalized new speaking tasks that require more extended spoken responses (i.e., respond to a situation and summarize a group discussion) and included an increased number of extended speaking and writing tasks. Additionally, it has revised the scoring criteria for productive skills. The updated scoring framework includes more detailed performance descriptors across discourse and linguistic dimensions such as content, development structure and coherence, and general linguistic range, and the automated scoring models were retrained to align with these revised criteria. As a result, the speaking and writing tasks examined in this study reflect an earlier version of the test, and accordingly, score- and task-based findings from the present study should be interpreted with caution.

Method

Participants

Thirty undergraduate international students enrolled in first-year writing (FYW) courses at a large public university in the United States participated in the study. As shown in Table 1, participants were 20.83 years old on average. Sixteen participants identified as female students, and 14 identified as male. They represented a wide range of linguistic backgrounds, including Hindi, Telugu, Vietnamese, Bengali, Gujarati, Arabic, and several other first languages. Academic majors included engineering and computer science, business, health and biology, and social sciences. The participants self-rated their English proficiency on a ten-point scale, with mean scores of 8.03 for speaking, 7.09 for writing, 7.93 for listening, and 7.93 for reading. They were enrolled in sections of the FYW course and volunteered to complete PTE Academic tasks, in-class writing tasks, and a post-task perception survey. While participants were enrolled in different sections, they received the same writing tasks and were evaluated using the same grading scales because the course curriculum was standardized across the program.

Table 1

Participant Characteristics

Characteristics		*N* = 30
Age	*Mean (SD)*	20.83 (2.59)
Gender	Female	16
	Male	14
First Language	Hindi	9
	Telugu	4
	Vietnamese	2
	Bengali	2
	Gujarati	2

	Arabic	2
	Other languages	9
Major	Engineering / Computer Science	13
	Business	6
	Health / Biology	4
	Social Sciences	3
	Other	4
Self-Rated Proficiency	0-10 scale: *Mean (SD)*	
	Speaking	8.03 (1.33)
	Writing	7.09 (1.19)
	Listening	7.93 (1.33)
	Reading	7.93 (1.33)

Instruments

PTE Academic

Participants completed one of the PTE Academic scored practice tests (Test E) using vouchers provided by Pearson. Although the practice test uses the same task formats and automated scoring framework as the operational PTE Academic test, it is administered outside of official test centers and does not involve the same level of standardized equipment checks or real-time technical monitoring. In this study, testing was supervised and took place in a computer lab, and participants used standardized equipment, including a university-provided computer and headset. From the full set of PTE Academic tasks, the present study focused on those that required test-takers to produce their own sentences or discourse. These included five speaking

tasks that elicited short monologic responses (describing images and re-telling a lecture) and one independent writing task that required an extended written response of approximately 200 to 300 words. Technical difficulties (recording failures or very low-quality recordings) for four participants contributed to PTE Academic Speaking scores of 10 (the minimum score out of 90). These Speaking scores were excluded from analyses. Similarly, two participants' PTE Academic writing scores were missing or excluded due to apparent technical issues (one participant did not receive a Writing score, while another received a Writing score of 10 but their retrieved independent writing task response seemed to be of much better quality). It should be emphasized that such technical issues are unlikely under standard operational testing conditions, which are conducted in controlled test centers with dedicated equipment and monitoring procedures.

Authentic Tasks

For the comparison of PTE Academic tasks and authentic academic work, students submitted three FYW writing tasks (letter, narrative, and analysis writing) and additionally completed one authentic speaking task (see Appendix A). For the letter writing task, students wrote a letter to the author of a published narrative, responding to the author's ideas with personal reflections. Five participants did not complete this assignment. The narrative writing task required students to build a three-page language-related narrative based on detailed memories. For the analysis writing task, they summarized a required reading and agreed or disagreed with its main points. These writing tasks were assigned and scored by the course instructors as part of their coursework, and the scores were shared with us with participants' informed consent.

In addition to the course-based writing tasks, participants completed an oral response task that was designed specifically for the purpose of this research project to approximate authentic

spoken interaction in FYW courses. The task required students to read a sample essay, presumably written by a peer, about environmental pollution, reflect on its content, organization, and language use, and discuss the essay with a teaching assistant for approximately 3-5 minutes. The research assistant guided the discussion with a set list of questions and followed-up as appropriate. This task was implemented in a standardized, consistent fashion for the purposes of this research project, intended to reflect authentic discussions of writing that occur in FYW courses but are typically not recorded or graded. The oral response task was scored using a modified analytic rubric that assessed pronunciation, fluency, vocabulary, grammar, and topic elaboration. Teaching assistant speech was trimmed prior to conducting linguistic analyses of the student speech.

Linguistic Measures

All spoken and written responses were analyzed for syntactic complexity, lexical complexity, accuracy, and fluency (CAF) using a set of established linguistic indices. The selection of linguistic features was guided by prior CAF research as core dimensions of productive language performance that are theoretically linked to task difficulty and proficiency development (Bulté & Housen, 2012; Wolfe-Quintero et al., 1998). These measures have been widely used in validation studies examining alignment between standardized test tasks and academic language use (e.g., Riazi, 2016; Brooks & Swain, 2014). The selected indices provide observable indicators of linguistic properties that are sensitive to L2 proficiency and task design. The CAF measures applied to both modalities include mean length of unit (AS-unit for speaking: MLAS; T-unit for writing: MLT), mean length of clause (MLC), clauses per unit (C/AS for speaking; C/T for writing), lexical density (proportion of content words), academic word frequency (AW Frequency), Measure of Textual Lexical Diversity (MTLD), Proportion of Error-

free units, errors per unit, errors per 100 words. Additionally, we analyzed words per minute (WPM) for the spoken responses and T-units per sentence (T/S), coordinate phrases per clause (CP/C), complex nominals per clause (CN/C), and dependents per Nominal for the written responses. Table 2 presents a summary of all linguistic measures used in this study.

These syntactic and lexical measures were extracted using automated natural language processing tools validated in prior work (i.e., Tool for the Automatic Analysis of Lexical Diversity (TAALED); Kyle et al., 2021; Tool for the Automatic Analysis of Lexical Sophistication (TAALES); Kyle & Crossley, 2015; Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC); Kyle, 2016; Lu, 2010). Accuracy indices were generated with the Auto Error Analyzer (Mizumoto, 2025). To ensure the validity of automated outputs, error annotations were manually reviewed, and for spoken performance, punctuation-based error counts were excluded. This combination of automated processing and targeted manual checking ensured that all CAF measures reflected genuine linguistic features of participant performance. Although the algorithms used to generate PTE Academic scores are not publicly available, prior research on automated scoring systems indicates that productive skill scores are derived from multidimensional models incorporating features related to fluency, lexical sophistication, grammatical accuracy, syntactic patterning, and intelligibility (Deane, 2013; Zechner & Evanini, 2020). The CAF measures examined in the present study, which were selected based on theoretical relevance and prior research, overlap conceptually with several of these dimensions, but they represent only a partial approximation of the full scoring models. That is, the selected linguistic measures do not intend to map directly onto individual scoring components but to provide an indirect, theory-driven lens for examining how selected linguistic properties of test responses correspond to those observed in authentic academic tasks.

It should also be noted that PTE Academic productive tasks scores are based on analytic rubrics that assess multiple traits at the discourse and linguistic levels (automated scoring systems were trained on human ratings using such rubrics). These rubric-based criteria differ from the evaluative practices used in first-year writing courses, which often prioritize rhetorical development, genre awareness, and revision processes over time. Consequently, differences in scoring frameworks and underlying constructs should be considered when interpreting alignment between PTE Academic scores and course-based performance.

Table 2

Linguistic Measures

Domain	Description	Modality
Fluency	Words per minute (WPM)	Speaking
	Total number of words	Writing
Syntactic	Mean length of unit	Both
Complexity	Mean length of clause	Both
	Clauses per unit	Both
	T-units per sentence	Writing
	Coordinate phrases per clause	Writing
	Complex nominals per clause	Writing
	Dependents per nominal	Writing
Lexical Complexity	Proportion of content words (CW Density)	Both
	COCA Academic word frequency log (AW Frequency)	Both
	Measure of Textual Lexical Diversity (MTLD)	Both
Accuracy	Errors per 100 words	Both
	Errors per unit	Both
	Proportion of error-free units	Both

Note. The term *unit* is used as a cover term for the following units of production: AS-unit (Analysis of Speech Unit, Foster et al., 2000) for speaking and T-unit (Terminal Unit) for writing (Bulté & Housen, 2012).

Task Perception Survey

Immediately after completing the PTE Academic test, participants responded to a task perception survey that included both 7-point Likert-scale and open-ended items (see Appendix B). The survey assessed cognitive demand, time management, task relevance, difficulty relative to coursework, instruction clarity, and enjoyment for both the PTE Speaking and PTE Writing tasks. The open-ended items elicited specific comments on features of the PTE tasks that made them easier or more difficult. While completing the survey, participants were also provided with a handout containing the PTE task prompts, which they could refer to as needed.

Analyses

Descriptive statistics and correlations were used to address the study's research questions. Due to the small sample size, restricted proficiency range of students, and potential relevance of modest correlations (e.g., those that might be considered "small", around .3, according to Plonsky & Oswald, 2014), we considered a more liberal *p*-value threshold of .10 in addition to the conventional .05 cutoff when evaluating statistical significance. For one variable that was majorly skewed (i.e., letter writing scores), Spearman's rank-order correlation was used to provide a more robust estimate of association. The decision to conduct pairwise correlational analyses, rather than multilevel modeling or aggregation across tasks, was motivated by both the limited sample size and the analytical goal of examining task-specific patterns. This approach allowed for a more transparent assessment of how different PTE Academic task types relate to authentic task performance.

Results

In this section, we present quantitative findings related to students' performance on the PTE Academic tasks, first-year writing (FYW) course tasks, and the oral response task, as well

as their perceptions of these tasks. In addressing the first research question, we examined students' performance on the PTE Academic and authentic tasks after excluding extreme low scores. Table 3 summarizes descriptive statistics for the remaining participants. PTE Speaking scores ranged from 30 to 87 (M = 57.77, SD = 13.52), and PTE Writing scores ranged from 37 to 80 (M = 58.75, SD = 10.56). The FYW writing tasks demonstrated relatively strong performance overall. Letter writing scores averaged 9.20 (full score = 10), while narrative and analysis writing averaged 85.03 and 87.83, respectively (full score = 100). Oral response scores ranged from 14 to 24 (M = 19.96, SD = 2.94; full score = 25). We note that the FYW writing tasks varied in difficulty, with the Letter Writing task presenting less of a challenge than other tasks, and that grades generally reflect a criterion-referenced orientation (i.e., mastery of learning objectives).

Table 3

Descriptive Statistics for Quality Scores by Task

Measure	*n*	*Mean*	*SD*	Median	Min	Max
PTE Speaking (/90)	26	57.77	13.52	55.00	30	87
PTE Writing (/90)	28	58.75	10.56	57.00	37	80
Oral Response (/25)	30	20.10	2.93	20.00	14	24
Letter Writing (/10)	25	9.20	1.37	10.00	5	10
Narrative Writing (/100)	30	85.03	9.05	83.50	70	99
Analysis Writing (/100)	30	87.83	8.83	89.50	72	100

PTE Speaking demonstrated statistically significant positive associations with oral response pronunciation (r = .46, p = .018) and fluency (r = .42, p = .031). The correlation between PTE Speaking and oral response total score was moderate (r = .32) but did not reach statistical significance. In contrast, PTE Writing scores did not show significant associations with any of the first-year writing task scores. Correlations with letter, narrative, and analysis writing were small and non-significant, suggesting limited score-level correspondence between PTE

Writing and course-based writing performance. To some extent, the high average grades for FYW course assignments and criterion-referenced marks can account for the weakness of these correlations.

Table 4

Correlations between PTE Tasks and Authentic Tasks

Tasks	*n*	*r*	*p*
Oral Response			
Pronunciation	26	**.458***	.018
Fluency	26	**.424***	.031
Vocabulary	26	.101	.623
Grammar	26	.188	.357
Topic Elaboration	26	–.126	.541
Total Score	26	.316	.116
FYW Tasks			
Letter Writing (Spearman's rho)	25	.001	.996
Narrative Writing	28	.210	.283
Analysis Writing	28	-.032	.872

Note. * = $p < .05$

Table 5 presents descriptive statistics for speaking CAF measures. Across the five PTE Speaking tasks, mean length of AS-unit (MLAS) values ranged from 13.47 to 17.03, while the oral response task had a lower value of 12.97. A similar pattern was found for mean length of clause (MLC), where PTE tasks averaged between 7.11 and 10.07, compared to 7.33 for the oral response. We note that the oral response task was interactive, and some shorter turns likely account for some of these length-related differences at the task level. Lexical measures also showed clear variability (e.g., MTLD values ranging from 29.56 to 48.41).

Table 5

Mean and Standard Deviations for Speaking CAF Measures by Task

Measure	PTE Image 1	PTE Image 2	PTE Image 3	PTE Lecture 1	PTE Lecture 2	Oral Response
Syntactic Complexity						
MLAS	15.67 (5.31)	13.47 (4.54)	16.25 (4.94)	17.03 (4.34)	14.42 (3.67)	12.97 (2.54)
MLC	9.79 (1.99)	8.26 (1.50)	9.93 (2.27)	10.07 (1.94)	7.11 (1.54)	7.33 (0.97)
C/AS	1.65 (0.66)	1.69 (0.79)	1.71 (0.62)	1.75 (0.58)	2.06 (0.47)	1.78 (0.30)
Lexical Complexity						
CW Density	0.56 (0.04)	0.58 (0.04)	0.59 (0.06)	0.60 (0.05)	0.67 (0.04)	0.57 (0.05)
AW Frequency	3.13 (0.16)	3.10 (0.17)	3.17 (0.15)	3.12 (0.17)	3.29 (0.08)	3.09 (0.27)
MTLD	31.72 (12.65)	29.56 (12.52)	38.77 (17.63)	38.41 (14.14)	48.41 (22.11)	38.06 (10.32)
Accuracy						
Error-free AS-unit Proportion	0.46 (0.27)	0.51 (0.27)	0.46 (0.28)	0.45 (0.25)	0.29 (0.18)	0.36 (0.20)
Errors per AS-unit	1.06 (0.76)	0.93 (0.74)	1.12 (0.91)	1.03 (0.68)	1.66 (0.75)	1.07 (0.68)
Errors per 100 words	5.62 (3.33)	4.96 (3.67)	6.00 (4.37)	5.55 (3.51)	10.14 (4.69)	6.82 (3.40)
Fluency						
Words/Min	110.53 (33.88)	125.32 (36.06)	118.09 (33.22)	118.29 (31.29)	103.34 (28.39)	101.17 (22.95)

Table 6 shows correlations between CAF measures in PTE tasks and corresponding measures in the oral response. Several syntactic and lexical measures exhibited positive associations. MLAS for one of the PTE image description tasks correlated significantly with oral response MLAS ($r = .38$, $p < .05$). Errors per AS-unit and errors per 100 words for one of the PTE lecture re-telling tasks correlated significantly with those of the oral response ($r = .42$ and $r = .45$, respectively, both $p < .05$). WPM for an PTE image description task showed a strong positive correlation with oral response WPM ($r = .62$, $p < .01$). These findings suggest that selected PTE Speaking tasks share linguistic demands with the authentic oral task, although the strength of associations varied across specific CAF dimensions.

Table 6

Correlations between PTE Speaking and Oral Response Task CAF Measures

CAF Measures	PTE Image 1	PTE Image 2	PTE Image 3	PTE Lecture 1	PTE Lecture 2
MLAS	.09	.15	**.38***	.23	**.34+**
MLC	**.36+**	.16	.11	.08	–.01
C/AS	–.08	.14	–.19	–.15	.03
CW Density	–.07	.03	–.10	–.10	.16
AW Frequency	–.07	–.15	.16	–.18	–.16
MTLD	–.07	–.03	.08	.19	.28
Prop. EF AS-units	**.33+**	.14	.11	–.17	–.10
Errors per AS-unit	.16	.03	–.05	.20	**.42***
Errors per 100W	**.44***	–.18	.17	.21	**.45***
WPM	**.62****	.10	–.03	.02	–.07

Note. + = $p < .10$; * = $p < .05$; ** = $p < .01$

Descriptive statistics for writing task CAF measures appear in Table 7. The FYW writing tasks displayed consistent syntactic and lexical complexity. For instance, MLC ranged from 6.97

to 7.89, and MTLD ranged from 34.53 to 46.00. These values tended to be comparable to those of the PTE Writing task.

Table 7

Mean and Standard Deviations for Writing CAF Measures by Task

Measure	PTE Writing	Letter Writing	Narrative Writing	Analysis Writing
Syntactic Complexity				
MLC	7.89 (1.65)	6.97 (1.37)	7.67 (1.28)	7.53 (1.42)
MLT	20.17 (4.17)	16.39 (3.27)	18.36 (3.24)	18.56 (3.24)
C/T	1.63 (0.33)	1.51 (0.28)	1.64 (0.33)	1.60 (0.26)
T/S	1.32 (0.15)	1.17 (0.19)	1.25 (0.18)	1.28 (0.16)
CP/C	0.26 (0.15)	0.22 (0.12)	0.23 (0.11)	0.23 (0.11)
CN/C	1.44 (0.33)	1.28 (0.27)	1.32 (0.29)	1.35 (0.29)
Dependents/N	2.38 (0.55)	1.60 (0.57)	1.68 (0.43)	1.61 (0.52)
Lexical Complexity				
CW Density	0.60 (0.05)	0.59 (0.05)	0.60 (0.04)	0.58 (0.04)
AW Frequency	3.07 (0.12)	3.05 (0.08)	3.08 (0.09)	3.06 (0.08)
MTLD	46.00 (15.66)	34.53 (12.29)	37.37 (12.24)	36.01 (12.75)
Accuracy				
Errors per 100 Words	13.42 (3.54)	9.66 (4.20)	7.61 (3.93)	7.83 (3.89)
Errors per T-unit	2.38 (0.56)	1.60 (0.57)	1.28 (0.48)	1.44 (0.49)
Prop. Error-free T-units	0.14 (0.12)	0.34 (0.18)	0.40 (0.23)	0.36 (0.20)
Fluency				
Total Words	203.10 (60.39)	130.03 (43.26)	225.70 (68.08)	233.20 (63.43)

Table 8 presents correlations between PTE Writing CAF measures and FYW CAF measures. Accuracy measures displayed the strongest and most consistent associations. Errors per T-unit correlated significantly with PTE Writing for letter, narrative, and analysis writing (*r*s ranged from .41 to .51, all $p < .05$). Proportion of error-free T-units for analysis writing also correlated significantly with the same measure for PTE Writing ($r = .43$, $p < .05$). Lexical sophistication (AW Frequency) for analysis writing showed a significant negative association with PTE Writing ($r = .42$, $p < .05$). These results indicate that accuracy, and to a lesser extent lexical sophistication, accounts for much of the shared variance between PTE Writing performance and FYW writing performance.

Table 8

Correlations between PTE Writing and First-Year Writing Task CAF Measures

CAF Measures	Letter Writing ($n = 25$)	Narrative Writing ($n = 28$)	Analysis Writing ($n = 28$)
MLC	.03	–.12	–.04
MLT	.04	–.01	.06
C/T	–.01	.06	.13
T/S	–.15	–.15	–.22
CP/C	–.05	–.16	.15
CN/C	–.04	–.10	–.05
Dependents/N	–.08	–.19	–.13
CW Density	–.07	.18	–.02
AW Frequency	.08	–.28	**–.42***
MTLD	.08	–.06	**.34+**
Errors per 100W	–.09	–.30	–.11
Errors per T-unit	**.41***	**.51****	**.42***
Prop. EF T-units	.23	.27	**.43***
Total Words	.19	.32	–.11

Note. + = $p < .10$; * = $p < .05$; ** = $p < .01$

The third research question involved international students' perceptions of the PTE Academic productive tasks. Table 9 presents perception ratings for PTE Speaking tasks. Students reported moderate cognitive demand ($M = 4.57$, $SD = 1.89$) and moderate difficulty relative to classroom tasks ($M = 4.53$, $SD = 1.57$). Instruction clarity received the highest ratings ($M = 6.17$, $SD = 1.15$). Enjoyment was also relatively high ($M = 5.43$, $SD = 1.48$). Qualitative comments, summarized in Table 10, revealed several recurring themes. The most frequently mentioned challenge was limited time for responding ($n = 15$), followed by difficulty with interpreting topics or visuals ($n = 6$), limited planning time ($n = 5$), difficult vocabulary ($n = 4$), and accents ($n = 2$). Four responses fell into miscellaneous categories. These themes demonstrate that time constraints and topic interpretation were central concerns when completing the PTE Speaking tasks.

Table 9

Descriptive Statistics for Speaking Perception Items

Measure	*Mean*	*SD*	Min	Max
Enough time for planning	4.60	1.87	1	7
Cognitive demand	4.57	1.89	1	7
Planning time spent	3.30	1.34	1	6
Relevance to current courses	4.93	1.31	3	7
Difficulty compared to class tasks	4.53	1.57	1	7
Instruction clarity	6.17	1.15	3	7
Enjoyment	5.43	1.48	2	7

Table 10

Speaking Task Features Leading to Ease or Difficulty (Qualitative Themes)

Theme	Count

Limited time for responding	15
Limited time for planning	5
Difficult topics / hard-to-interpret visuals	6
Difficult vocabulary	4
Accents	2
Other / unspecified features	4

Note. Totals exceed the number of participants because some respondents selected multiple features.

Table 11 summarizes perception ratings for PTE Writing tasks. Students reported adequate time for planning, writing, and revising ($M = 5.20$, $SD = 1.73$), and relatively high instruction clarity ($M = 6.37$, $SD = 1.07$). Overall enjoyment was similar to speaking ($M = 5.70$, $SD = 1.56$), and cognitive demand was moderate ($M = 4.37$, $SD = 1.92$). Qualitative responses in Table 12 highlighted several challenges. Students noted limited time for brainstorming ($n = 3$) and limited contextualization of writing prompts ($n = 3$). A small number of responses indicated that successful completion required substantial reading ($n = 1$). Other comments addressed favorable task features such as adequate writing time ($n = 5$) and relevant or interesting topics ($n = 4$). These findings suggest that although students generally perceived the PTE Writing tasks positively, some aspects of task design constrained their ability to plan or contextualize their responses.

Table 11

Descriptive Statistics for Writing Perception Items

Measure	*Mean*	*SD*	Min	Max
Enough time for planning, writing, revising	5.20	1.73	1	7
Cognitive demand	4.37	1.92	1	7
Planning time spent	3.93	1.82	1	7
Relevance to current courses	5.43	1.41	2	7

Difficulty compared to class tasks	4.27	1.36	2	7
Instruction clarity	6.37	1.07	3	7
Enjoyment	5.70	1.56	2	7

Table 12

Writing Task Features Leading to Ease or Difficulty (Qualitative Themes)

Theme	Count
Limited time for brainstorming	3
Limited contextualization of tasks	3
Reading skills needed / heavy reading load	1
Enough time for writing	5
Relevant, interesting topics	4

Discussion

This study examined how international students' performance on PTE Academic productive tasks corresponds to their performance on authentic academic tasks and how they perceive the cognitive demand and relevance of these tasks. Taken together, the results offer a multi-faceted view of the cognitive validity of the PTE Academic Speaking and Writing sections. PTE Academic Speaking scores showed moderate associations with the pronunciation and fluency dimensions of the authentic oral response task. These findings suggest that the global PTE Speaking score captures delivery-related components of academic oral interaction, which might have arisen from the fact that both PTE and oral response tasks require largely spontaneous language production under time pressure, with limited opportunity for planning. At the same time, the absence of a statistically significant association with the overall oral response score indicates that score-level alignment is partial and dimension-specific rather than uniform.

In contrast, PTE Writing scores did not exhibit meaningful associations with performance on the first-year writing tasks. This pattern suggests that, at the level of holistic scores, PTE

Writing may capture aspects of writing proficiency that are not strongly reflected in authentic writing outcomes. Course-based academic writing typically involves an extended process that includes planning, drafting, revision, and access to external resources, whereas the PTE Writing task requires time-constrained, single-draft production with no external support. These differences in composing conditions might have resulted in weak score correspondence between standardized writing and authentic academic tasks.

It is also plausible that limited score alignment may also reflect differences in construct emphasis and measurement properties. PTE Writing scores are based on analytic evaluation of multiple traits such as content relevance or coherence, whereas course-based writing tasks often prioritize genre-specific rhetorical development, engagement with readings, and responsiveness to feedback over time. The present study focused primarily on selected linguistic features and did not directly target higher-level discourse organization or content-related dimensions that contribute to holistic writing scores. In addition, although the FYW writing tasks are authentic representations of course requirements, they may function less effectively as measurement instruments for differentiating writing proficiency, particularly when score distributions are restricted. In our study, this issue is particularly salient for the letter writing task, which demonstrated limited variability and reduced discriminative capacity. The FYW writing task grades also reflect a criterion-referenced approach to assessment in which mastery of learning objectives is emphasized, rather than distinguishing relative levels of performance. This highlights a challenge of using instructor grades from authentic tasks in research that examines support for the extrapolation of test scores to performance in the academic domain.

Consistent with this interpretation, Llosa and Malone (2019) reported that scores from the TOEFL iBT independent task had weaker associations with course-based writing performance

than those from the integrated writing task, which approximates academic writing assignments more closely. They further found that TOEFL iBT writing scores were stronger predictors of grades on first drafts than final drafts, illustrating the impact of feedback and revision on the link between test and non-test writing performance which we may have observed in the present study. In particular, the FYW writing task grades in this study are final grades given to the final drafts of scaffolded, process-based writing instruction. Taken together, these considerations suggest that the observed weak score correspondence reflects an interaction of composing conditions, construct coverage, and the measurement characteristics of the authentic tasks, rather than a simple absence of relationship.

CAF analyses revealed partial linguistic alignment between PTE Academic tasks and authentic tasks, with accuracy and fluency measures demonstrating the most consistent correspondence across modalities. For speaking, several PTE tasks displayed linguistic properties that were comparable to the oral response, although patterns varied across measures. Mean length of AS-unit values for PTE image description and oral response tasks were similar, and a significant within-person correlation emerged between these measures. Fluency also aligned well for some tasks, particularly PTE image description, where WPM strongly correlated with the oral response WPM. Error-based accuracy measures showed additional points of alignment, most notably errors per AS-unit and errors per 100 words. These findings suggest that specific PTE speaking tasks elicit linguistic features that resemble authentic oral performance in the academic context. The lack of consistency across PTE tasks may have to do with prompt effects (e.g., specific lexical and syntactic demands for describing a particular graph) or the short duration of each task. Future research might profit from aggregated analysis of each student's set of speaking performances.

For writing, CAF measures revealed stronger and more consistent alignment. Accuracy measures were particularly informative. Errors per T-unit for all three FYW tasks showed significant associations with the corresponding measure for PTE Writing. The proportion of error-free T-units for analysis writing also correlated significantly with PTE Writing. These findings indicate that the accuracy of written language production in PTE Academic corresponds meaningfully to accuracy in course-based writing, supporting the argument that PTE Writing captures core grammatical and discourse-level abilities required in academic contexts. On the contrary, lexical sophistication for analysis writing showed a significant negative relationship with that of PTE Writing, a pattern potentially suggesting that higher-level lexical choices in FYW tasks may not be fully mirrored in the PTE Academic essay. Overall, the CAF findings suggest that linguistic alignment between test tasks and authentic tasks is strongest for accuracy measures and somewhat weaker for syntactic and lexical complexity measures, especially in speaking. Syntactic and lexical complexity of writing products may be more subject to specific task demands and topics, as different communicative purposes and topics require distinct linguistic forms (Biber & Conrad, 2009), while underlying linguistic competence that impacts accuracy is more broadly at play in any writing task.

In addition, international students studying in the U.S. generally viewed PTE Academic tasks as moderately demanding yet fair, although they highlighted several areas of difficulty related to time constraints and task interpretation. Student perceptions offered an additional source of validity evidence. Ratings for both PTE Academic Speaking and Writing indicated moderate cognitive demand and task difficulty. Students generally judged the tasks as relevant to their academic courses and viewed instructions as clear. These results suggest that the tasks were perceived as accessible and fair, even if challenging.

The qualitative responses provided deeper insight into the sources of difficulty. For speaking, the dominant concerns involved time constraints for responding and planning, as well as challenges in interpreting visual prompts and unfamiliar vocabulary. These comments align with the cognitive validity framework, in which time pressure and conceptual load can constrain attentional resources for language formulation. For writing, students noted limited contextualization of prompts and limited time for brainstorming, along with some positive comments regarding topic interest and adequate time for writing. These perceptions suggest that the PTE Writing task is viewed as relevant but somewhat constrained relative to the more elaborated and genre-specific tasks encountered in FYW courses.

Conclusion

The findings provide converging evidence for the partial alignment between PTE Academic productive tasks and authentic speaking and writing tasks. Score-based associations were modest but meaningful for speaking, while CAF analyses revealed stronger alignment for accuracy and fluency measures across both modalities. Student perceptions indicated that the tasks were reasonably representative of academic demands, although certain design features posed challenges related to time constraints and task interpretation. The findings provide qualified support for the interpretation of PTE Academic Speaking scores as indicators of certain aspects of academic speaking performance, particularly those related to pronunciation and fluency. Evidence for alignment in writing was more limited, suggesting that extending writing score interpretations to contexts emphasizing extended academic writing processes requires some caution (Llosa & Malone, 2019).

Several limitations should be acknowledged. First, although FYW tasks and the oral response task were designed to reflect authentic academic activities, they were not independently

validated as instruments for differentiating writing or speaking proficiency. In particular, we used actual instructor grades given for the FYW writing tasks which reflect a criterion-referenced approach to assessment and do not afford norm-referenced interpretations. As a result, the scoring of these authentic tasks may not have provided sufficient measurement sensitivity and/or range of scores necessary to support robust, easily interpretable correlational analyses. This limitation does not reflect a deficiency in the pedagogical value of the tasks but highlights the distinction between authenticity and measurement precision in validation research. Second, as noted earlier, the data were collected prior to the implementation of recent updates to PTE Academic. As a result, direct generalization of task-level findings to the current operational version of PTE Academic should be made with caution.

With regard to the current study designs, the sample size was small and drawn from a single institution, which limits generalizability. The oral response task was relatively short and may not fully represent the range of academic speaking genres encountered in university contexts. PTE Academic tasks were administered as a practice version, which was scored using an automated system different from that being currently used for the operational test. Additionally, although CAF measures provide a rich profile of linguistic performance, automated extraction can introduce noise in the case of syntactically complex or disfluent speech. We also note that the selection of CAF measures in this study, while motivated by theory, represent only a small selection of measures that have been explored in research and that are used in the PTEA automated scoring engines.

Future research could examine more or different CAF features, consdier multi-draft authentic writing tasks (with a focus on initial/first drafts), or include extended speaking performances to investigate linguistic alignment with PTE Academic over longer spans of

discourse. Studies involving multiple institutions and larger samples would further strengthen claims regarding cognitive validity. Finally, qualitative interviews or stimulated recall protocols could yield deeper insights into students' cognitive processing during test tasks.

References

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

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge University Press.

Biber, D., Reppen, R., & Staples, S. (2017). Exploring the relationship between TOEFL iBT scores and disciplinary writing performance. *TESOL Quarterly, 51*, 948-960.

Brooks, L., & Swain, M. (2014). Contextualizing performances: Comparing performances during TOEFL iBT and real-life academic speaking activities. *Language Assessment Quarterly, 11*, 353-373.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Investigating complexity, accuracy and fluency in SLA* (pp. 21-46). Amsterdam: John Benjamins.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics, 19*, 254-272.

Crossley, S. A., & McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing, 26*, 66-79.

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., & James, M. (2006). *Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL*. Princeton, NJ: Educational Testing Service.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing, 18*, 7-24.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*(3), 354–375.

Kane, M. (2013). Validating the interpretation and uses of test scores. *Journal of Educational Measurement, 50*, 1-73.

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication* (Doctoral Dissertation). Retrieved from http://scholarworks.gsu.edu/alesl_diss/35.

Kyle, K. & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly, 49*(4), 757-786.

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the validity of lexical diversity using direct judgements. *Language Assessment Quarterly, 18*(2), 154-170.

Llosa, L., & Malone, M. E. (2019). Comparability of students' writing performance on TOEFL iBT and in required university writing courses. *Language Testing*, *36*, 235-263.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*(4), 474-496.

Mizumoto, A. (2025). Automated analysis of common errors in L2 learner production: Prototype web application development. *Studies in Second Language Acquisition, 47*, 867-884.

Pearson. (2021). Score guide for test takers: PTE Academic (Version 15 – April 2021). London, UK: Pearson Education.

Plonsky, L., & Oswald, F. L. (2014). How Big Is "Big"? Interpreting Effect Sizes in L2 Research: Effect Sizes in L2 Research. *Language Learning*, *64*(4), 878–912.

Riazi, M. (2013). Concurrent and predictive validity of Pearson Test of English Academic (PTE Academic). *Papers in Language Testing and Assessment, 2*, 1-27.

Riazi, M. (2016). Comparing writing performance in TOEFL-iBT and academic assignments: An exploration of textual features. *Assessing Writing, 28*, 15-27.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics, 22*, 27-57.

Wagner, E. (2020). Duolingo English test, revised version July 2019. *Language Assessment Quarterly, 17*, 300-315.

Weigle, S. C. (2002). *Assessing writing.* Cambridge: Cambridge University Press.

Weigle, S. C. (2011). *Validation of automated scores of TOEFL iBT tasks against non-test indicators of uniting ability* (Research Report TOEFL iBT-15). Princeton, NJ: Educational Testing Service.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity*. University of Hawaii at Manoa, Second Language Teaching and Curriculum Center.

Zechner, K., & Evanini, K. (Eds.). (2020.). *Automated speaking assessment: Using language technologies to score spontaneous speech*. Routledge.

Appendix A

Oral Response Task

Reflecting and Giving Feedback on Writing

In this task, you will read an essay written by a student on the topic of pollution and the environment. As you read the essay, you should reflect on its content, organization, and language use. Take notes on your thoughts about the essay, focusing on its strengths and weaknesses. You may take notes on this paper. Afterward, you will be asked to share your thoughts on the essay and recommendations for the author.

Essay Prompt:

Every country in the world has problems with pollution and damage to the environment. Do you think these problems can be solved?

Student Essay:

I think that my country has problems with pollution to the environment like all other countries. This problem is normal for Russia. We have big problems with transport because there are too much cars in our country. And because of that we have problems with atmospeer, air in my city and in all Russia is really dirty and sometimes I can't make a sigh because it smells around me and of course around that cars on the road. I've heard about tradition of one country. They don't go anywhere by car one day a month or a year, they just use bycicle or their feet. I think it could be very good if we had a tradition like that.

So, what about the rivers and the seas? Yeah, there are some really good and clean rivers and seas where you can go, but there are not many of them. Once I saw the river OB in my city, it was about two years ago but I stil remember that in some places it was not blue, it was green or purple I didn't really understand because it had different colors.

I don't know what should we do. Maybe we should just open our eyes and look what we did. But Russian people don't care about the world around them many people care only about themselves and that's all.

So, the best idea is look around and try to do something good for our planet and for us and our children.

Letter Writing Task

Purpose:

This assignment requires students to write a 250-400 word letter or email to the author of a narrative essay. The purpose is to communicate a personal response to the essay, demonstrate comprehension of the author's ideas, and explain relevant personal experiences.

Task Description:

Students select a published narrative essay (excluding Amy Tan's "Mother Tongue") and write a letter expressing their thoughts about the essay. The letter should refer to the author's main points, explain why particular ideas were meaningful or confusing, and connect the essay to the student's own experiences. Students may also include questions for the author.

Submission Requirements:

- 250-400 words
- Times New Roman, 12-point font
- Double-spaced with 1-inch margins
- MLA formatting

Narrative Writing Task

Purpose:

This assignment requires students to write a language-related narrative that integrates detailed memories with reflective analysis.

Task Description:

Students compose an approximately 1,000-word narrative describing an important language learning event or a person who influenced their language development. The essay should include vivid sensory details, dialogue when appropriate, and a clear explanation of the significance of the chosen memory or individual.

Submission Requirements:

- 3-4 full pages (approximately 1,000 words)
- Times New Roman, 12-point font
- Double-spaced with 1-inch margins
- MLA formatting

Analysis Writing Task

Purpose:

This assignment requires students to write a 500-750 word analytical essay that summarizes a required reading and responds to its key ideas using personal experience.

Task Description:

Students summarize the major points of a selected reading and then support or critique each point using concrete personal examples. The essay must incorporate at least one signal phrase, one parenthetical citation, one paraphrase, and one direct quotation. The essay must conclude with a correctly formatted MLA Works Cited entry.

Submission Requirements:

- 500-750 words
- Times New Roman, 12-point font
- Double-spaced with 1-inch margins
- MLA formatting

Appendix B

TEST PERCEPTION QUESTIONNAIRE

Please answer the following questions for the **speaking** and **writing** tasks of PTE Academic, referring to the test task prompts as needed.

Speaking Tasks

1. Did you feel you had enough time to plan your ideas before responding? *(Circle one)*
   Strongly disagree **1 – 2 – 3 – 4 – 5 – 6 – 7** Strongly agree

2. How mentally demanding were the speaking tasks? *(Circle one)*
   Not at all demanding **1 – 2 – 3 – 4 – 5 – 6 – 7** Very demanding

3. How much time did you spend planning your response before speaking? *(Circle one)*
   None **1 – 2 – 3 – 4 – 5 – 6 – 7** A lot

4. How relevant were the speaking tasks to real-world or academic speaking situations you encounter? *(Circle one)*
   Not at all relevant **1 – 2 – 3 – 4 – 5 – 6 – 7** Highly relevant

5. How difficult were the speaking tasks compared to other speaking activities you have done in class? *(Circle one)*
   Much easier **1 – 2 – 3 – 4 – 5 – 6 – 7** Much harder

6. What specific aspects of the speaking tasks made it easy or difficult for you? *(Open-ended)*

7. How clear were the instructions for the speaking tasks? *(Circle one)*
   Not at all clear **1 – 2 – 3 – 4 – 5 – 6 – 7** Very clear

8. Did you enjoy completing the speaking tasks? *(Circle one)*
   Not at all **1 – 2 – 3 – 4 – 5 – 6 – 7** Very much

9. Do you think the speaking tasks accurately reflected your speaking abilities? Why or why not? *(Open-ended)*

10. Were there any aspects of the speaking tasks that felt unfamiliar, confusing, or unexpected? *(Open-ended)*

Writing Tasks

11. Did you feel you had enough time to plan, write, and revise your response? *(Circle one)*
    Strongly disagree **1 – 2 – 3 – 4 – 5 – 6 – 7** Strongly agree

12. How mentally demanding were the writing tasks? *(Circle one)*
    Not at all demanding **1 – 2 – 3 – 4 – 5 – 6 – 7** Very demanding

13. How much time did you spend planning your response before writing? *(Circle one)*
    None **1 – 2 – 3 – 4 – 5 – 6 – 7** A lot

14. How relevant were the writing tasks to real-world or academic writing situations you encounter? *(Circle one)*
    Not at all relevant **1 – 2 – 3 – 4 – 5 – 6 – 7** Highly relevant

15. How difficult were the writing tasks compared to other writing activities you have done in class? *(Circle one)*
    Much easier **1 – 2 – 3 – 4 – 5 – 6 – 7** Much harder

16. What specific aspects of the writing tasks made it easy or difficult for you? *(Open-ended)*

17. How clear were the instructions for the writing tasks? *(Circle one)*
    Not at all clear **1 – 2 – 3 – 4 – 5 – 6 – 7** Very clear

18. Did you enjoy completing the writing tasks? *(Circle one)*
    Not at all **1 – 2 – 3 – 4 – 5 – 6 – 7** Very much

19. Do you think the writing tasks accurately reflected your writing abilities? Why or why not? *(Open-ended)*

20. Were there any aspects of the writing tasks that felt unfamiliar, confusing, or unexpected? *(Open-ended)*