ENGLISH PROFICIENCY TESTING: HOW ARE WE DOING?

Kym Taylor, PhD Janna Schaeffer, PhD

International Testing Services Center (ITSC-Group)

National College Testing Association
July 2025

SESSION AGENDA

INSPIRATION FOR PROJECT

STUDY DESIGN, IMPLEMENTATION, AND FINDINGS

IMPLICATIONS FOR TESTING

ADDITIONAL APPLICATIONS OF FINDINGS

GROUP ACTIVITY

FINAL THOUGHTS

INSPIRATION FOR PROJECT

BACKGROUND

OVERVIEW OF PROFICIENCY TESTING IN US

CURRENT RESEARCH UNDERSTANDING

APPROACH TO IDENTIFYING WHAT IS MISSING

Major Proficiency Testing Companies in the US

TEST	CREATOR	RELEASE	TEST VERSIONS
TOEFL	Educational Testing Service	1964	TOEFL iBT (academic)* TOEFL Essentials (academic and general)
TOEIC	Educational Testing Service	1979	Listening and Reading (workplace)* Speaking and Writing (workplace) Bridge (workplace)
IELTS	British Council	1989	IELTS Academic* IELTS General Training IELTS Life Skills
PTE	Pearson	2009	PTE Academic* PTE Core (general/workplace) PTE Home (visa and immigration)
DET	Duolingo	2016	Duolingo English Test

BACKGROUND

- More than one million international students in the US (Israel & Batalova, 2021)
  - bring culture, innovation, and revenue to higher education institutions (Hegarty, 2014).
- English language proficiency tests a rite of passage (Abedi & Sanchez, 2021)
- Proficiency testing industry continuing to grow (Baer & Martel, 2020; Institute of International Education [IIE], 2018)
- Field advancing daily to better fit needs of test takers and institutions (MacGregor, Yen, & Yu, 2022).

BACKGROUND

We know that…

- proficiency tests have a certain amount of predictive power (Walyuo & Panmei, 2021).

But research (and anecdotal evidence) also suggests that…

- testing scores only minimally correlate with student performance once on campus (Ihlenfeldt & Rios, 2023).

BACKGROUND

- Proficiency measurement not a simple task due to many factors, including:
  - Imbalance of skill proficiency (Koizumi et al., 2022; Longabach & Peyton, 2018; Rubio & Hacking, 2018)
  - Diversity of test taker population (e.g., first language, educational background) (Yoo et al., 2019)
  - Equity (Sabbaghan & Fazel, 2023)
  - Scoring consistency within and across assessments (Kang, Rubin, & Kermad, 2019; Li et al., 2022)
  - Accessibility (Dewi et al., 2023)
  - Growing acceptance of World Englishes (McNamara, 2023; Monfared, 2020)
  - Security and cheating concerns (Clark, 2023; Kim, 2022; Niu et al., 2024)

GAP IN RESEARCH

- Limited research on stakeholder perceptions of the alignment of proficiency tests with actual ability
- Absence of research exploring decision-makers' perceptions of proficiency test fit in specific contexts

RESEARCH QUESTIONS

Research Questions:

Which tests are post-secondary institutions relying on for English proficiency assessment?

In what ways are these assessments meeting the needs of their institutions?

In what ways are assessments lacking?

What would be on decision makers' wish lists?

STUDY APPROACH

One way to assess the current functionality of proficiency tests is through comprehensive **needs analysis** (i.e., needs assessment) (Munby, 1978; Richterich & Chancerel, 1980).

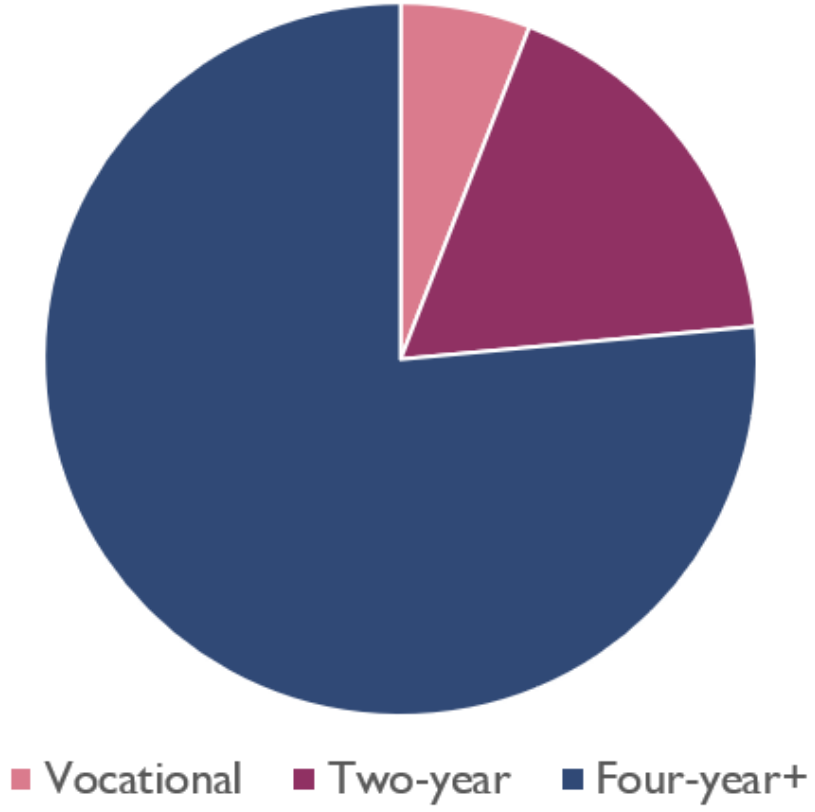
Needs analysis can:

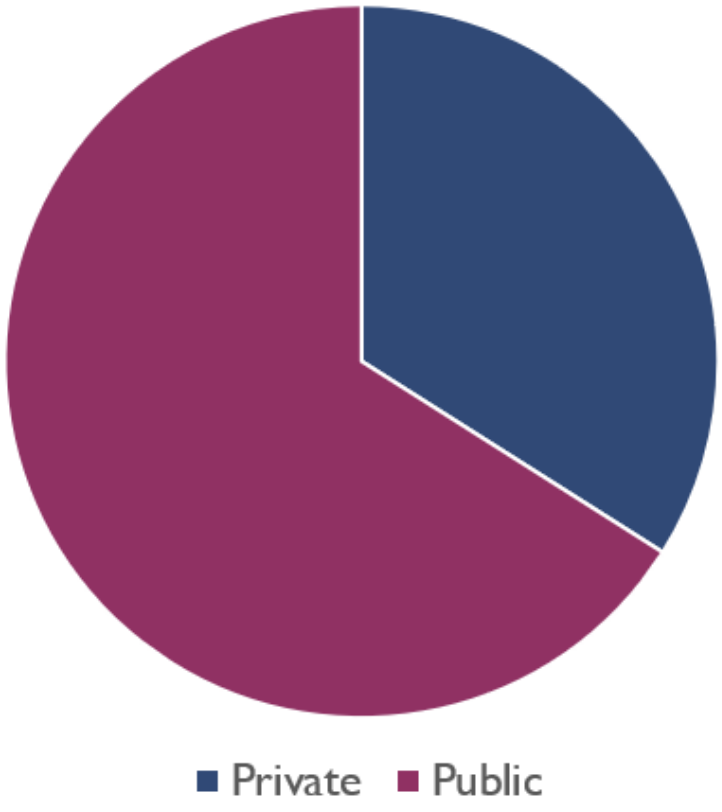
- Inform decision-making and assist in strategic planning (Peykari et al., 2013)
- Help optimize resources (Huber et al., 2015)
- Tailor solutions to problems in specific contexts (Allen, 2015)
- Engage stakeholders, thus building relationships (Han et al., 2020)
- Provide a way to monitor progress (Laurent et al., 2023)

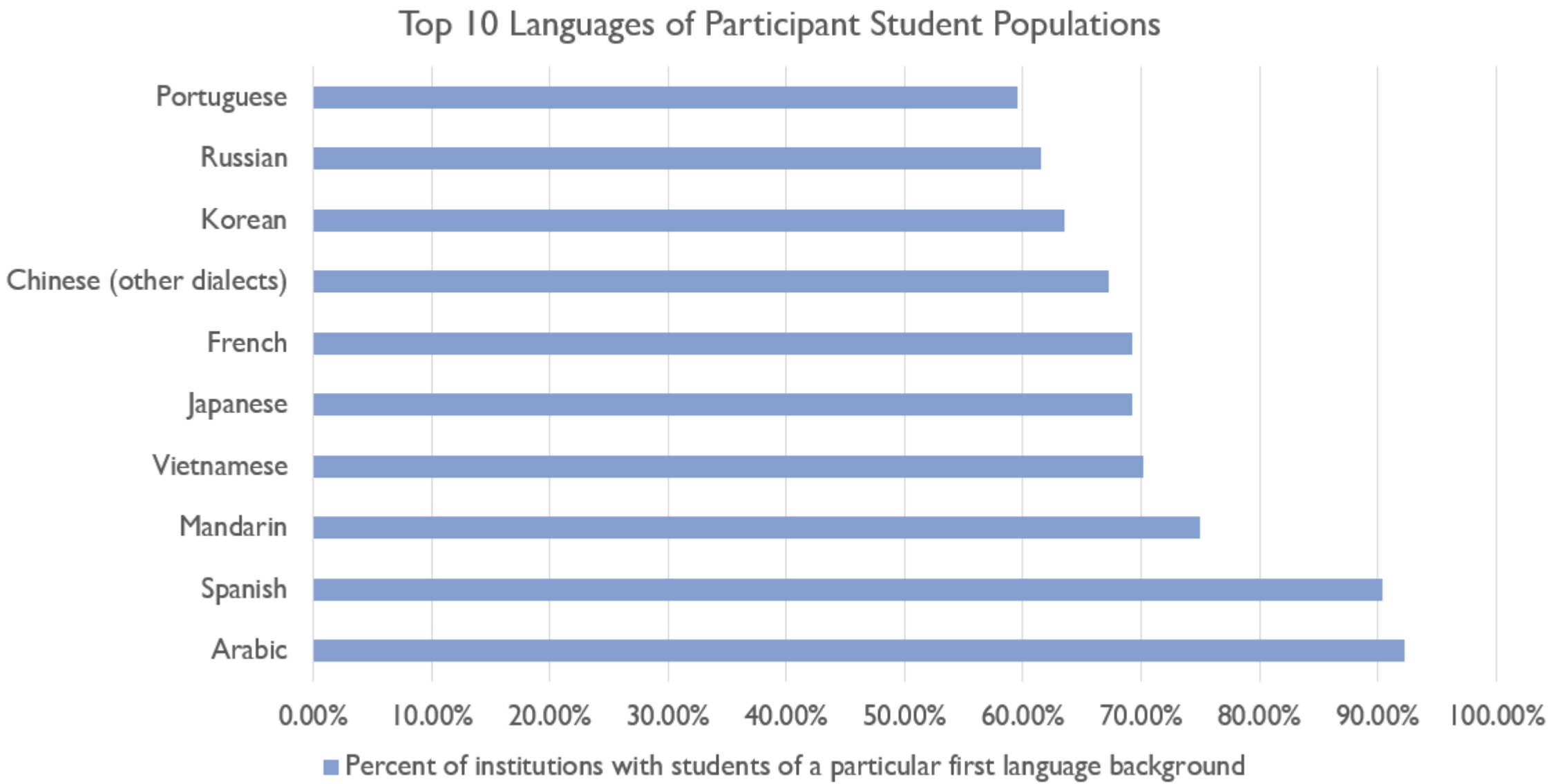
PARTICIPANTS

- 50 decision-makers at institutions across the US
- ESL/EAP program administrators at postsecondary institutions
- Graduate degrees in TESOL / Applied Linguistics / Education
- Mean time in field = 24 years

PARTICIPANT AFFILIATION





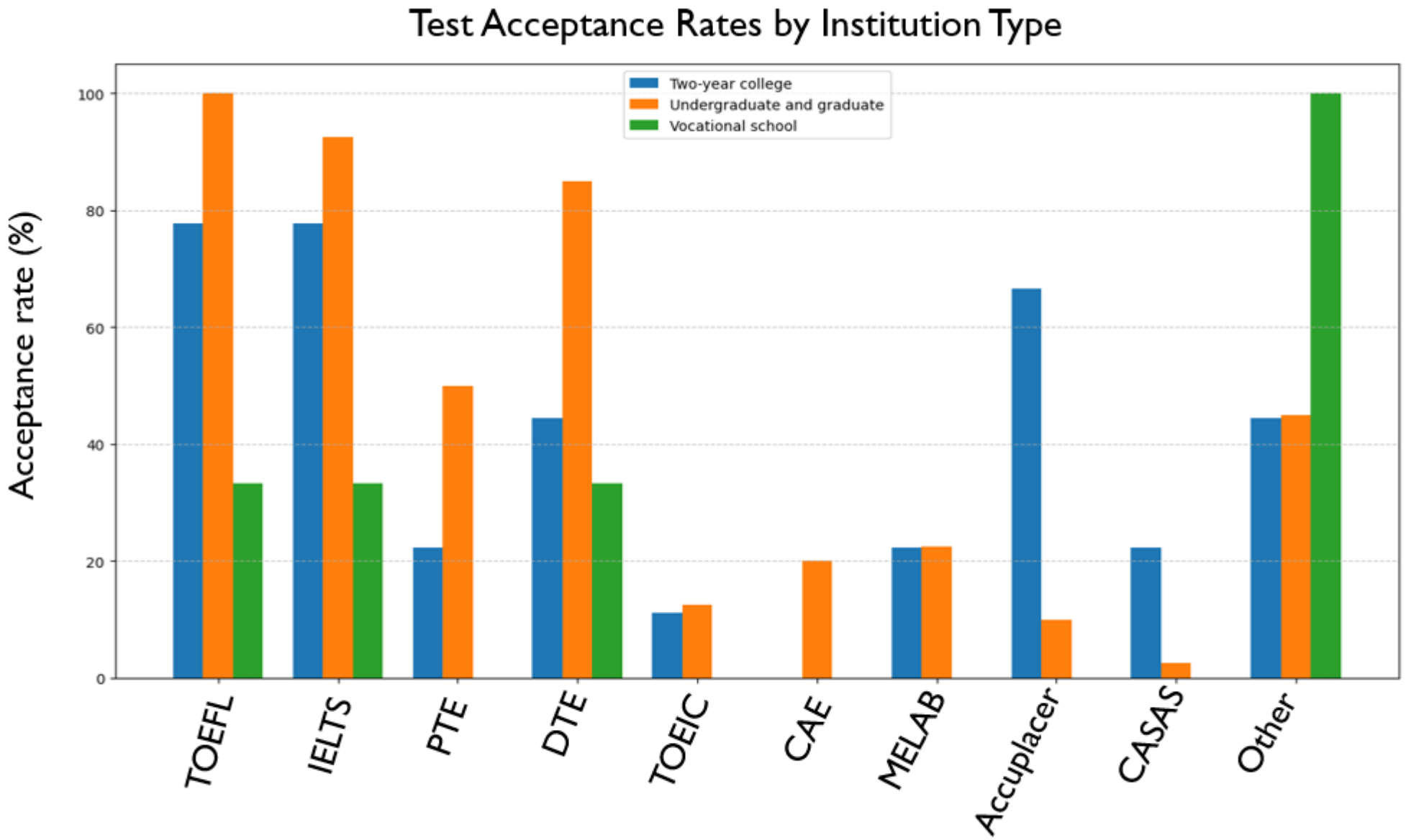

Top 10 Languages of Participant Student Populations
Portuguese
Russian
Korean
Chinese (other dialects)
French
Japanese
Vietnamese
Mandarin
Spanish
Arabic
0.00%
10.00%
20.00%
30.00%
40.00%
50.00%
60.00%
70.00%
80.00%
90.00%
100.00%
Percent of institutions with students of a particular first language background

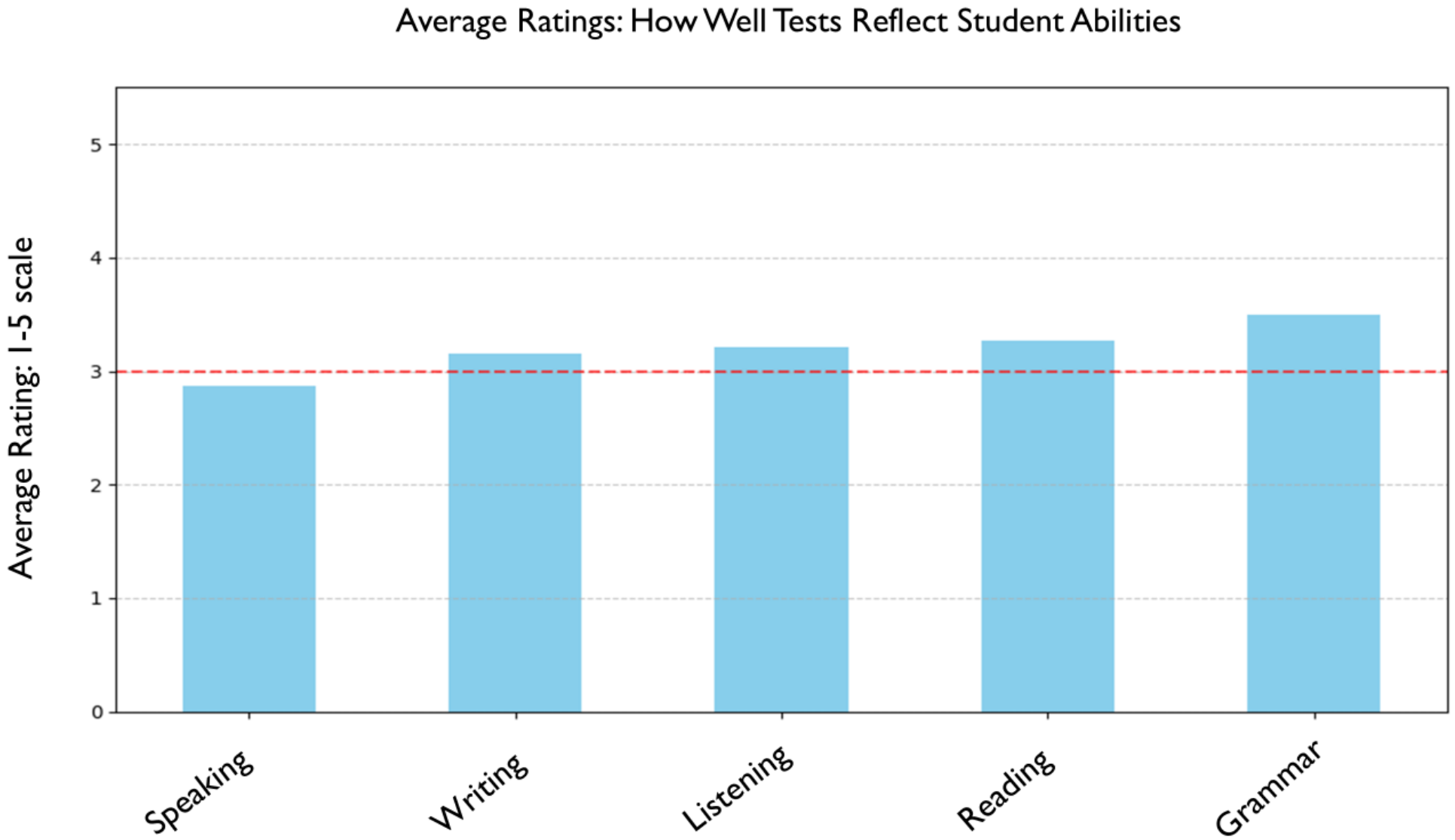
METHOD

COMPARATIVE, MIXED-METHODS APPROACH

ONLINE SURVEY CREATED USING LIMESURVEY 6.5

QUANTITATIVE FINDINGS


Test Acceptance Rates by Institution Type
Two-year college
Undergraduate and graduate
Vocational school
Acceptance rate (%)
100
80
60
40
20
0
TOEFL
IELTS
PTE
DTE
TOEIC
CAE
MELAB
Accuplacer
CASAS
Other


Average Ratings: How Well Tests Reflect Student Abilities
Average Rating: 1-5 scale
0
1
2
3
4
5
Speaking
Writing
Listening
Reading
Grammar

READING task ranking by importance (Top 10)

Ranking	Task
1	Summarizing information
2	Inferential questions about a nonfiction text
3	Literal questions about a nonfiction text
4	Inferential questions about a fictional text
5	Retelling
6	Vocabulary questions based on text content
7	Connecting ideas
8	Making predictions
9	Literal questions about a fictional text
10	Annotation

LISTENING task ranking by importance (Top 10)

Ranking	Task
1	Literal questions about formal audio input (e.g., lecture, speech)
2	Oral or written responses
3	Retelling
4	Literal questions about casual audio (e.g., everyday conversation)
5	Notetaking
6	Inferential questions about formal audio input (e.g., lecture, speech)
7	Inferential questions about casual audio (e.g., everyday conversation)
8	Matching exercises
9	Tasks that pair audio and visual information
10	Gap-fill exercises

WRITING task ranking by importance (Top 10)	
Ranking	**Task**
1	Writing prompt (suggestion of a general writing topic)
2	Summary of a reading text
3	Open response to a reading text
4	Summary of reading and listening texts combined
5	Open response based on reading and listening texts combined
6	Opinion based on reading and listening texts combined
7	Response to a fixed question
8	Opinion based on a reading text
9	Opinion based on an audio clip
10	Sentence starters

SPEAKING task ranking by importance (Top 10)	
Ranking	**Task**
1	Composition (as evidenced in a student's production of original speaking material)
2	Indirect (embedded in assessment of other skill areas)
3	Spontaneous conversation
4	Problem solving aloud based on an issue posed
5	Talking about themselves in relation to a specific topic (e.g., hobbies, traditions, beliefs)
6	Describing a chart or infographic
7	Making an oral presentation
8	Responding aloud to a reading text
9	Sharing an opinion about a particular topic
10	Describing a photo or picture

QUALITATIVE ANALYSIS

- Qualitative data were analyzed through iterative coding (Glaser & Strauss, 1967) and reflexive thematic analysis (Braun & Clarke, 2006).
- Some categories (e.g., skill type) were inherent; others (e.g., grounds for acceptance of a particular test) emerged from the coding process.

KEY QUALITATIVE FINDINGS: Grammar

- Grammar is often integrated rather than tested separately.
- Current tests reliably assess grammar skills, but through different approaches.
- Grammar tests have design limitations.
- Grammar is intertwined with other skills.
- Cultural and educational backgrounds influence grammar knowledge and acquisition.
- Grammar testing is becoming gradually less emphasized in academic contexts.

KEY QUALITATIVE FINDINGS: Grammar

"Generally reliable though sometimes students are prepped so well they're better at testing grammar than authentic grammar." (P37)

"I think most tests of grammar are unfair, biased, or difficult for students to understand. I'm not convinced that this score is very helpful." (P27)

"I am of the school that grammar should be integrated and not isolated. When we communicate as global English speakers, it is the message and the interaction that matters, not the verb tense." (P8)

KEY QUALITATIVE FINDINGS: Reading

- Current tests are reasonably effective for reading assessment.
- There are limitations to current tests' predictive accuracy regarding reading skills.
- Test design matters: Skill integration and authentic reading texts are key.
- Volume and complexity of reading tasks may be too much for even proficient students.
- Current tests may need supplementation of on-site assessment for assessing reading.

KEY QUALITATIVE FINDINGS: Reading

"Tests do not cover the range of reading genres needed for academic success." (P2)

"The reading portion of TOEFL is especially challenging for many students. If they can do well on it, it seems to be a good predictor of their future ability to read academic texts." (P3)

"The reading tests measure the comprehension well; however, students can be on the lower end of the score in this skill. Inferencing for comprehension and application of the material can be difficult to measure compared to academic courses." (P36)

KEY QUALITATIVE FINDINGS: Listening

- Major tests seen as most reliable for listening assessment, but with limitations.
- Many tests fail to reflect real-world or academic settings in listening tasks.
- Test design matters: Time constraints, non-repetition of material, varied accents, and test anxiety particularly affect listening assessment.
- Vocabulary knowledge and topic familiarity are key to listening tasks.
- Listening skills often fail to correlate with other skills.
- Noticeable gaps exist between listening test scores and demonstrated ability on campus.

KEY QUALITATIVE FINDINGS: Listening

"The listening sections of these tests tend not to be natural and often are testing for vocabulary comprehension.They also tend to be the shortest portions of the tests.They are still useful, but I wish there was more to check ability to understand fluent English." (P24)

"Test don't prepare students for the duration and intensity of listening skill necessary for a day of 3 or 4 lectures in multiple topics." (P6)

"Listening skills are not reflected accurately because they do not reflect real-life experiences with access to more than one- time listening opportunities." (P43)

KEY QUALITATIVE FINDINGS: Writing

- The two most popular tests are seen as the most reliable in assessing writing.
- The scope of writing tests may be too narrow; key skills are overlooked.
- Test design and delivery matters: Time constraints, lack of training, and access to AI all affect equitable writing assessment.
- Writing scores are often misaligned with on-campus performance, especially considering summer decline.
- Standardized proficiency tests should be supplemented with in-house measures.
- Writing skills are often part of an imbalanced profile (productive vs. receptive skills).

KEY QUALITATIVE FINDINGS: Writing

"Academic writing skills are not adequately tested on these tests because test-takers memorize the key parts of a traditional five paragraph essay and only practice for that. However, after the test, these test taking skills do not effectively transfer to their overall ability to write in academic setting." (P43)

"It is my opinion that the key issue for writing proficiency assessment is careful multiple assessments, carefully designed grading criteria, and carefully selected, clearly and simply described, writing instructions." (P1)

"[The most popular tests] are consistent predictors of student academic performance. I like that they each have two writing tasks which require synthesis, critical thinking, and summarizing." (P42)

KEY QUALITATIVE FINDINGS: Speaking

- Major tests are seen as most reliable for speaking assessment.
- Major limitations include failure to provide spontaneous, interactive opportunities, failure to account for test taker anxiety, and cultural comfort levels.
- Conversational proficiency does not equal academic proficiency.
- Face-to-face interviews are considered best, but they are costly and time-consuming.
- It is important to supplement pre-admission testing with on-campus assessment.

KEY QUALITATIVE FINDINGS: Speaking

"The tests do a good job showing the students' ability to speak conversationally. Many of them still need quite a bit of work on presentation skills and [contributing to] group work." (P3)

"Not all tests reflect students' actual speaking skills because they are assessed different ways and measured differently." (P43)

"[The major tests] are reliable, valid, and useful [for speaking assessment] but the time and cost required for their administration and review cannot be [underestimated] when budgeting for these tests and their use in admissions decision making." (P1)

KEY QUALITATIVE FINDINGS: Vocabulary

- Vocabulary and topic familiarity are closely linked.
- Task complexity affects student performance.
- It is difficult to disentangle assessment of vocabulary from comprehension.

KEY QUALITATIVE FINDINGS: Vocabulary

"Some students have a limited vocabulary based on topic familiarity; those two go hand-in-hand. And if the task is complex, they don't always understand the directions well--or everything they need to do. They only do some of the task." (P3)

"Grammar (and vocabulary) testing demands that the test include a wide range of items, in line with corpus research on the grammatical features of academic English in the institutional context you have in mind." (P1)

QUALITATIVE FINDINGS: What needs improvement?

- Close gap between standardized testing content and real-world language use.
- Decrease emphasis on test-taking strategies vs. true proficiency.
- Control for inaccurate assessment due to limitations in test-taker vocabulary and topic familiarity.
- Raise awareness of how test design influences performance.
- Reconsider impact of time constraints, test-taker anxiety, and testing conditions.
- Consider role of scoring by humans vs. AI: How do we demonstrate accuracy and alignment?
- Sharpen attention to potential issues of fairness and bias.
- Reduce mismatch between scores and actual ability upon arrival.

KEY TAKEAWAYS, IMPLICATIONS, AND DISCUSSION

- Current tests adequately assess proficiency, but there is plenty of room for improvement.
- Tests must be affordable, accessible, and broad in scope while adhering to reasonable time limits.
- Decision-makers support the field of assessment's shift toward integrated content.
- Test designers need to take stakeholder concerns into consideration.
- There is room for partnership between testing companies and institutions for streamlined pre-and post-admission assessment.

LIMITATIONS AND FUTURE DIRECTIONS

Limitations

- Number of participants
- Balance of institution type (vocational, 2-yr., 4-yr.+)
- Selection bias

Future directions

- Study replication with different stakeholder perspectives (e.g., test takers)
- Further investigation of correlation between proficiency scores and performance; predictive capabilities
- Investigating test acceptance trends over time

QUESTIONS?

NEEDS ANALYSIS APPLICATIONS

ACTIVITY

FINAL THOUGHTS

THANK YOU!