THE ROUTLEDGE HANDBOOK OF LANGUAGE PROGRAM DEVELOPMENT AND ADMINISTRATION

Edited by Alan V. Brown, Cori Crane,
Beatrice C. Dupuy, and Estela Ene

First published 2025

ISBN: 9781032420240 (hbk)
ISBN: 9781032421001 (pbk)
ISBN: 9781003361213 (ebk)

Chapter 11

Understanding the Role of Standardized Exams in Second Language Programs

DOI: 10.4324/9781003361213-14

The funder for this chapter is Brigham Young University.

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

11

UNDERSTANDING THE ROLE OF STANDARDIZED EXAMS IN SECOND LANGUAGE PROGRAMS

Troy L. Cox, Alan V. Brown, and Margaret E. Malone

Standardized language tests are used for many purposes in pK-12 and higher education, including program admission and placement, certification of language attainment and proficiency, and graduation requirements. Despite the many standardized language tests that exist, many administrators and instructors possess limited understanding of how they are designed, what functions they fulfill, and how to interpret and use the results. The influx of AI in the language testing space has only exacerbated this problem. Key program stakeholders often rely on these tests to make pressure-filled and high-stakes decisions but may be doing so without understanding their limitations.

However, the issue is not limited to the tests and their immediate results; often the test design and philosophies that undergird the test construct—i.e., what the test purports to measure—influence what is taught in language classes. This principle is frequently referred to as *washback*, or the influence tests and their administration can have on curriculum design and classroom pedagogy (Hughes, 2003). The design of the tests whose scores an institution accepts communicates to potential students—whether intentionally or not—what language skills the program values among its students. Students will prepare arduously for standardized tests—especially admissions tests, take practice exams, and strive for passing scores without necessarily learning language skills in a way that will help them survive and thrive both within the academic setting and in everyday life. This principle applies not only to language tests used for admission but those used at any stage in our programs, e.g., placement, proficiency, graduation, or teacher certification. The challenge language programs face is choosing standardized assessments appropriate for their local context and intended use. The greater the stakes, the more examinees are incentivized to attain a specific score, potentially at the expense of acquiring useful language skills. The selection of tests that include language from the context(s) outside the instructional setting—i.e., the target language use (TLU) domain—is in the best interest of the examinee and the institution for long-term success (Bachman & Palmer, 2010). Thus, the tests we adopt as administrators and how those tests conceptualize second language ability ought to reflect our institution's mission and values.

There are many reasons standardized language exams have had and will continue to have a far-reaching impact on higher education. First, standardized tests can be a cost-effective and

DOI: 10.4324/9781003361213-14

efficient way for institutions to reduce their budgetary expenses by offloading measurement costs to students. These test fees could be a required component of enrollment in a particular program of study, such as an admissions or placement test, in which the prospective student pays to certify they have the requisite language skills needed to succeed in the program (i.e., admission) or to start at a certain level (i.e., placement). Other tests (e.g. the College Board Advanced Placement exams or the College-level Examination Program tests) that allow students to obtain course credit from previously gained language skills without needing to enroll in classes can be an efficient and economical way for students to certify their language ability. Second, the highly technical nature of item construction and the extensive validation process requiring large numbers of students instills confidence in the validity and reliability of the exams. Test development and revision require resources and expertise that most programs do not have and could not undertake independently. Third, for an exam administered to thousands of students across an entire country or region, comparisons across students and programs are easily made. Indeed, standardized language exams are rather ubiquitous in modern post-secondary education; however, their development and resulting scores remain a mystery to many stakeholders.

In this chapter we aim to demystify standardized exams by describing how they are developed; what they are used for; what impact they have on curricula, learners, and instructors. We begin by briefly discussing the history of standardized language tests and making the case for their use in language programs. We also will describe a small set of standardized exams commonly used in post-secondary language programs, reviewing relevant research pertaining to their usefulness in making accurate inferences of candidates' abilities, i.e., validity argument. Finally, we offer recommendations and cautions on how they might be (mis)used at the local level.

Fundamental Considerations

A Brief History of Standardized Testing

Standardized testing has a long history, originating in imperial China as a solution to cronyism in civil service appointments. A standardized testing approach was developed to ensure that job candidates were selected based on merit rather than family background. These first tests involved evaluating a small, controlled sample of tasks to predict future performance in uncontrolled situations (Wainer, Bradlow, & Wang, 2007). The advent of standardized job-screening tests revolutionized employment by prioritizing what individuals knew over whom they knew.

In the United States, language testing in the late 19th and early 20th centuries focused primarily on ancient languages like Latin and Greek. These tests emphasized grammatical and lexical knowledge, assessing what examinees knew about the language rather than their ability to use it in conversation (Barnwell, 1996). However, following World War I, there was a shift toward teaching modern, spoken languages that coincided with the US military's need to assess soldiers' skills in a variety of domains, including foreign language, as the US began emerging as a world superpower. This led to the development of standardized testing protocols to objectively evaluate the abilities of individuals from diverse educational backgrounds (Spolsky, 2000).

Simultaneously, the College Board, in collaboration with American colleges and universities, began creating standardized tests across various disciplines. These tests aimed to allow high-achieving secondary students to earn college credit and to assist colleges in making admission decisions. Larger trends in the educational landscape, coupled with the growing interest in modern languages and practical language proficiency, influenced language testing. Language tests, many of which consisted of multiple-choice questions, were increasingly used for university admission, placement into language programs, course credit, program completion, and professional certification.

The emergence of psychometrics—the scientific measurement of psychological constructs—further strengthened the role of standardized tests. Testing companies began providing technical reports with concrete evidence of their tests' validity and reliability, thereby enhancing confidence in their assessments (Spolsky, 2000). Institutions increasingly outsourced test development to these professional companies, which employed experts in assessment, statistics, and psychometrics. While this third-party approach promoted uniformity across educational institutions, it also required careful consideration to ensure the appropriate use of test results in local contexts. For example, a listening test featuring only American English might be suitable for students planning to study in the United States but might be less effective for those attending universities in the British Isles or Australia.

Standardized tests have become deeply embedded in educational systems, influencing various aspects of education from admissions to professional certification. Despite their widespread use, these tests must, first, be understood insofar as their development is concerned, and, second, be carefully evaluated to ensure that they meet the specific needs of the institutions and individuals they serve.

Standardized Test Development

Test development generally includes four phases: design, operationalization, administration, and reporting (Bachman & Palmer, 2010). The design phase is crucial, as it delineates not only the content but also how the subsequent phases will align with the initial specifications. This phase begins with a language needs assessment which evaluates what language learners need to do and results in a test framework and item specifications (Davidson & Lynch, 2008). Standardized tests are designed to assess specific content, skills, or both as determined by the test developers and must be administered under standard conditions.

The design process starts by identifying the test's purpose, the language to be tested, and the assertions that can be drawn from the results, such as determining if a test taker has sufficient language skills for an academic program in an English-medium university (Bachman & Palmer, 2010). The test framework defines all aspects of the test: purpose, intended examinee population, outcomes, language characteristics (Davidson & Lynch, 2008), item types (e.g., multiple choice, constructed response), scoring systems, time limits, and reporting of results.

Item specifications detail each item's characteristics, including difficulty level, content, and expected responses. This includes defining reading passage lengths, the number of distractors in multiple-choice items, and the prompts for constructed responses. These specifications ensure that the test faithfully measures the intended skills and knowledge. Two primary approaches to test development are norm-referenced and criterion-referenced tests, each with distinct methods for evaluating examinee performance.

Norm-Referenced Tests

Norm-referenced testing compares examinees against each other, ranking them from the highest to the lowest ability. The *norm* in norm-referenced refers to the statistical normal distribution, emphasizing an examinee's position within that distribution (Gaertner, 2022). Developers identify the population to be tested (e.g., heritage learners, K-5, ESL 1.5) and define the construct and target language use domain (e.g., reading textbooks at the university, interpreting in the medical field, conversing with locals as a researcher or volunteer, writing technical reports, etc.).

Developers then create test and items specifications with the aim to disperse examinee scores, typically by designing items that 30% to 70% of the examinee population can answer correctly (Carr, 2011). Tests designed with these parameters typically result in a normally distributed bell curve that can be equated with previous test administrations. Examinees can then receive a standard score, such as a transformed z-score, showing their position relative to others. This approach helps make fine distinctions, especially when space or resources are limited. For instance, if an international internship program has only five spots for 100 applicants, a norm-referenced test would identify the top 5%. Similarly, in cases where remediation funds are limited, the test can identify the lowest performers most in need of assistance.

The strength of norm-referenced testing is that institutions can set their own cut scores based on available resources, which makes this approach common for admissions tests. However, it can be challenging to infer what students can or cannot do based on external standards, as the items may not explicitly test those standards. Consequently, even top examinees might lack necessary skills for success in certain domains, while lower performers might have sufficient skills. Due to these limitations, many language tests rely on criterion-referenced approaches instead.

Criterion-Referenced Tests

While not as pervasive as norm-referenced testing, criterion-referenced testing is increasingly common, particularly for certifying language proficiency. This approach compares examinee performance against external standards, aiming to report how well they have mastered those standards. The *criterion* refers to these external standards, scales, or criteria linked to specific skills, which may be rooted in a curriculum or a hypothesized hierarchy of levels (Brown & Abeywickrama, 2018; Hughes, 2003).

Designing criterion-referenced tests starts with defining the population to be tested and identifying the appropriate criteria. Test specifications aim to fully represent these criteria, regardless of item difficulty. For example, if the criterion is that examinees can ask questions to learn personal information about a peer, this would be included in the test specifications and would appear at a certain level on the scoring scale. An item based on this criterion would be included even if nearly all examinees could answer it correctly. Conversely, if a criterion states that examinees can infer a speaker's unstated position in a debate, an item would be included even if only a tiny fraction could answer it correctly. The results are reported as a percentage of correct responses relative to the set standard, without considering the performance of other candidates.

While test developers can create their own criteria for criterion-referenced tests, the items can also be constructed to reflect external standards. School districts, for example, use these exams to determine how many students have mastered particular learning outcomes, while professional organizations use them to certify that examinees possess

the necessary skills for specific work environments. The strength of this approach lies in its direct relationship between results and standards, allowing for diagnostic information and detailed profiles of strengths and weaknesses. However, the test's usefulness is closely tied to the soundness of the criteria; if the criteria are flawed or misaligned with natural language development or the desired learning outcomes, the results may be difficult to interpret.

Criterion-Referenced Proficiency Scales in Language Testing

Among the more widely known scales used with criterion-referenced language testing are the Common European Framework of Reference (CEFR) and American Council on the Teaching Foreign Languages (ACTFL) Proficiency Guidelines. Both stem from the US government's Interagency Language Roundtable (ILR) scale which was created as functional descriptors of what civil servants could and could *not* do in the language, as opposed to what they knew about the language and could state in their first language (Lowe, 1988). Derived from the ILR in the early 1980s, the ACTFL Guidelines were adapted to the educational and commercial landscape and revised in 1986, 1999, 2012, and 2024. ACTFL has maintained and revised the descriptors and interpretations of them for official testing purposes in the academic context since the 1980s (Liskin-Gasparro, 2025). The CEFR was established by the Council of Europe to describe achievement in learning outcomes (Verhelst et al., 2009). Initially designed to be curricular in nature, teachers, schools, and institutions have broad latitude in defining how the guidelines are operationalized in their individual contexts—thus two different tests may operationalize the same CEFR level in very different ways. Because the CEFR has no regulatory board and tests tend to be developed locally for each language, test score users should exercise caution in their interpretation of the results and the accuracy of accompanying inferences about language ability.

Current and Common Standardized Language Exams and their Use

Test of English as a Foreign Language (TOEFL)

The Test of English as a Foreign Language (TOEFL), created by Educational Testing Service (ETS), is one of the most widely recognized standardized language tests. It is often a critical component of the admissions process for non-native English speakers seeking entrance into English-medium institutions for undergraduate or graduate studies. While the TOEFL score is only one part of an applicant's portfolio, it often carries significant weight, with many institutions setting a required minimum score for admission. The consequences of not meeting TOEFL score requirements highlight its high-stakes nature, as applicants failing to reach the minimum score are often denied admission. Additionally, there are concerns that even those who meet the score requirements may struggle with daily communication in an English-speaking environment. This high-stakes nature has fueled a large market for test preparation resources.

The TOEFL, particularly the internet-based version (TOEFL iBT), was designed to assess a learner's proficiency in academic English used in university settings. It tests four language skills: reading, listening, speaking, and writing. Speaking and writing responses are evaluated by both human and AI raters, while reading and listening sections are machine-scored. Each section is scored on a scale from 0 to 30, with a total possible score of 120 points (ETS, 2024).

Given the significance of the TOEFL in academic admissions, extensive research has been conducted to evaluate its predictive validity, that is, the candidate's ability to perform typical language tasks required of university students in academic situations, but not necessarily social encounters. Studies have explored the TOEFL's ability to predict candidates' GPA and graduation rates and found weak to moderate positive correlations. The same held true for the correlation between TOEFL score and the degree to which students felt well-prepared and coped effectively with the linguistic demands of university (Harsch, Ushioda, & Ladroue, 2016; Ihlenfeldt & Rios, 2023). Other research found significant positive correlations between TOEFL scores and required performance in specific coursework (Biber, Reppen & Staples, 2017; Llosa & Malone, 2019; Yeom & Llosa, 2024). The development of the TOEFL iBT itself was the result of over a decade of research into what constitutes academic language across various tasks (Biber & Gray, 2013; Chapelle, Enright & Jamieson, 2011).

While the TOEFL remains the most widely accepted English language admissions test in the US, other exams, such as the International English Language Testing System (IELTS), are also used internationally. IELTS has gained popularity, and research has found it to be a meaningful predictor of academic success (Gagen & Faetz, 2024; Schoepp, 2018). Program directors can utilize existing research to better interpret admission test scores and support students aiming to achieve the necessary scores on standardized English language tests.

ACTFL Language Proficiency Assessments

English language tests are not the only foci for higher education; world language tests are also important for language programs. ACTFL, through its exclusive licensee Language Testing International, offers a suite of standardized proficiency exams assessing speaking, listening, reading, and writing in 100 languages. These exams use a 10-point proficiency scale (e.g., Novice Low, Mid, High up to Superior), aligned with ACTFL's well-defined construct of language proficiency (Liskin-Gasparro, 2003).

The origins of these proficiency tests date back to the mid-20th century, following World War II and the Korean War. The US government recognized the need for a standardized oral proficiency protocol and organized the Interagency Language Roundtable (ILR) to address the issue. The ILR created a formal oral interview structure accompanied by an 11-point oral proficiency scale that has been used by the Foreign Service Institute (Liskin-Gasparro, 2003) ever since. ACTFL adapted this scale for academic contexts in the early 1980s, creating a 10-point scale with finer distinctions at lower levels. The ACTFL Oral Proficiency Interview (OPI) closely follows the ILR interview structure and originally mirrored the ILR level descriptions (Barnwell, 1996).

The OPI is a 20–30-minute structured interview that includes an introduction, warm-up, level checks, probes, and a wind-down. The level checks identify the candidate's lowest performance level, while the probes push until linguistic breakdown occurs. A wind-down ensures candidates finish with tasks they can comfortably complete. If any of these elements are missing, or if the candidate's highest and lowest abilities are not demonstrated, the speech sample is not ratable, and no official score can be given. All OPIs are conducted by ACTFL-certified testers and are second-rated by another certified tester to ensure reliability. Despite its somewhat open-ended nature, the ACTFL OPI is the most widely used oral proficiency test in the US outside of government contexts.

Since its introduction in the mid-1980s, some scholars have questioned the validity of the ACTFL OPI, citing concerns about the contrived nature of the interview and the relationship

between the proficiency guidelines and rating procedures (Bachman, 1988; Barnwell, 1996; Johnson, 2001; Lantolf & Frawley, 1988; van Lier, 1989). However, subsequent research has demonstrated that OPI ratings are reliable and useful for differentiating oral ability levels (Alpine, 2020; Dandonoli & Henning, 1990; Halleck, 1996; Surface & Dierdorff, 2003; Thompson, 1995, 1996). The proficiency guidelines have since been adapted for reading, listening, and writing, and are used in many US states as a certification requirement for language teachers with many setting a minimum score of Advanced Low.

Other world language exams created by testing companies, national governments, and professional teachers' organizations are also available: Standards-based Measure of Proficiency (STAMP) from Avant Assessment, *Diploma del Español como Lengua Extranjera* (DELE) from the Cervantes Institute in Spain, the National German Exam from the American Association of Teachers of German (AATG), the *Diplôme d'Etudes en Langue Française* (DELF), the Chinese Proficiency Test (HSK), and the Test of Proficiency in Korean (TOPIK).

Implications and Applications for Language Program Development and Administration

In this section we address three primary implications of standardized tests for language programs. First, we unpack the terms *standard* and *standardized* and offer recommendations for how to maximize the usefulness of standardized test scores for individual programs by aligning scores to the program's curricular content and level structure. Second, we discuss the implications of using test score comparison tables. Finally, we examine the impact of high-stakes standardized testing on students and their exam preparation, as well as on the curricula and programs that use them.

Standardized Exams and their (Mis-)Use in Language Programs

Inherent in the word *standardized* are two assumptions: (1) a standard of some sort has been identified—in our case one related to language and one related to candidate ability and identity, and (2) a deliberate, methodical process was undertaken to determine that standard, hence the suffix *-ized*. When a test developer purports to employ a specific standard, such as the ACTFL Proficiency Guidelines, they must be able to prove that they are using that standard accurately. For example, asserting that a non-ACTFL organization is issuing official scores on the ACTFL Guidelines is misleading.

Language program administrators would do well to ask themselves the following fundamental question: When is a bad test better than no test at all? We offer a few cautions from our own experiences when selecting standardized tests. First, decide what problem a standardized test will solve. Will it be to admit students into a program or to certify proficiency at the end of a course of study? It is difficult for a test to serve multiple purposes, so clearly understanding what it was originally designed to measure as compared to what some might claim the test measures is crucial. Using a test to serve a function for which it was not designed is the educational equivalent of using a drug prescription for a sickness other than the one the doctor diagnosed. The drug might treat some of the symptoms, but since it was never tested with a population suffering that particular malady, unforeseen side effects may arise. The same holds true for standardized tests that are not used as directed. Second, there should be a strong positive correlation between (1) the stakes of a test for students in the program and (2) the due diligence and care administrators use

in selecting exams. For instance, the stakes are lower in accurately placing local students into a program than determining whether to admit international students who will need to relocate to a new country. Additionally, we feel it is in the best interest of examinees and the institution for long-term success when tests accurately reflect the target language use domain (Bachman & Palmer, 2010). Table 11.1 lists some additional questions language program administrators and other decision-makers in language programs should ask before selecting or deciding to use a standardized test and some examples of what to consider for each question. A bellwether for determining the bona fides of test developers is the extent to which they adhere to the ILTA Code of Ethics (International Language Testing Association, n.d.)

Table 11.1 Considerations for Standardized Test Selection.

Primary Question	*Additional Points to Consider*
• What is the purpose of the test?	• Is it for admissions? Placement? Achievement? Course credit? (Note: It is often difficult for a test to serve more than one purpose with the same level of accuracy.)
• What does the test claim to measure?	• How do the test developers define what they are measuring (for example, for a writing test, what genre is being measured)? • Is there enough detail in the documentation to create practice items for examinees?
• How are users supposed to interpret test scores?	• Are there scoring guidelines for the different subsections? • How are productive skills (e.g., writing or speaking) scored? • How can users interpret what the scores mean in their program? • Have standard errors of measurement been provided?
• Does the test have any published research establishing a credible reliability and validity argument?	• Where is the research published (for example, a reputable journal or a paper on the test developer's website)? • How detailed is the technical report? • What are the qualifications of the authors? • Have the developers followed the ILTA code of guidelines?
• For what population has the test been developed?	• Does this population match yours (for example, using a test developed for middle school and high school students for university level students)?
• What language varieties are included in the test?	• Does the test rate results on a single variety of the language or do the developers include multiple reviewers from across varieties of the language in test development?
• Is the test appropriate for heritage/community learners?	• Does the test reflect the different populations of your institution?
• How is the test administered?	• If computer based, do you have the technical requirements to administer it? • Is it available in the geographic regions where the students are located?
• Which standards does it use?	• Do the standards match your goals for student outcomes? • Does the test developer have the authority to make the claims it does about the standards it uses?
• How much does the test cost?	• Who will pay for the test (student or institution)? Does the cost represent a financial burden for stakeholders?

Test Score Comparison Tables and their (Mis-)Use

To make it easier for stakeholders to make comparisons among tests, many test publishers will provide test score comparison tables. We offer caution with the use of these tables, even when they claim to measure the same construct. Each organization may and often does operationalize the construct differently, making comparisons unreliable. For instance, a writing score from one test might reflect a brief response to a picture prompt designed to elicit Basic Interpersonal Communication Skills (BICS), while another test might include lengthy responses to multiple prompts completed over an hour or more to assess academic writing and Cognitive Academic Language Proficiency (CALP) (see Cummins, 1979). Thus, comparing writing scores from these tests would be problematic due to fundamental differences in task specifications and as members of different genres that require unique language skills.

This issue is especially relevant in the context of test score comparison tables, which are often provided by test developers to show score concordances between different exams. These tables allow institutions to decide whether scores from various tests can fulfill the same requirement. However, such concordances require that the tests (1) measure closely related constructs, (2) have high correlations, (3) are administered under similar conditions, and (4) maintain similar levels of reliability across the test and its subsections (Dorans, 2004; Elliott et al., 2021).

Knoch and Fan (2024) evaluated the concordance practices of major English language test providers and found significant shortcomings. These providers often present vague or insufficient information on their websites, and the research behind their concordance tables frequently fails to meet best practice standards, with issues like inadequate sample sizes and omission of crucial data, such as subsection scores and standard error measurements. These deficiencies are concerning as they can impact high-stakes decisions affecting test takers.

The lack of rigor in concordance studies may stem from developers using convenience samples from existing databases instead of collecting data specifically for this purpose. The absence of regulatory oversight and minimal demand for high-quality work from test users exacerbate the problem. While concordance tables can be useful, they provide only one piece of evidence regarding the validity of score comparisons. Program directors should exercise caution when interpreting these relationships. For example, if a test claims to produce results comparable to those of a different test, like the TOEFL, directors should scrutinize the accompanying research to determine whether the claims are substantiated.

Impact of High-Stakes Standardized Testing

Tests that align with a program's goals can positively influence, or create *washback*, in language programs. Research on the ACTFL Assessment of Performance toward Proficiency in Languages (AAPPL) (Vyn, 2024) and other ACTFL assessments, including integrated performance assessments (Cubillos, 2010; Martel, 2019), demonstrates positive washback effects on teaching and learning. However, language students often have less experience with standardized language tests than with other kinds of assessments. Therefore, it is crucial that test takers are familiar with the testing process to ensure that their scores accurately reflect their abilities rather than their unfamiliarity with the test format. This preparation is especially vital for high-stakes exams.

Messick (1982) identified three potential outcomes from coaching students in test preparation:

1. **Enhanced Test-Taking Skills**: Coaching can improve students' executive functioning skills, such as time management, strategic guessing, and how to approach various question types common in language tests (Foster, Paulk & Dastoor, 1999). Familiarity with test logistics, like using recording devices or electronic keyboarding, is also crucial. Increased familiarity can lead to higher test scores that better reflect the students' true abilities because they are well-prepared for the test format and item types.
2. **Genuine Improvements in Abilities**: Coaching may provide more opportunities for language use, leading to genuine improvements in the skills being tested. As students practice, they solidify foundational material and apply the language in future situations, resulting in higher test scores that accurately reflect their true abilities (Shanks, Don, Boustani & Yang, 2023).
3. **Superficial Score Increases**: The third outcome is a concern, as coaching might teach shortcuts that enable students to choose or produce the correct answer without the necessary knowledge or skills. This results in inflated test scores that do not accurately reflect the students' abilities. This disconnect is more likely when test tasks do not align with real-world tasks.

Figure 11.1 provides a graphic representation of the three aforementioned outcomes of test preparation and coaching identified by Messick.

In summary, test preparation can help students improve if it focuses on developing language skills rather than merely test-taking strategies. Language programs should choose tests that align with their educational philosophy and notions of language competence, as students—not surprisingly—tend to tailor their practice to the test for which they are preparing.

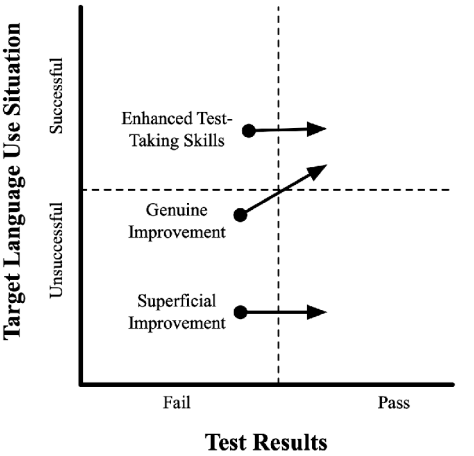

Figure 11.1 Impact of Coaching on Language Skills and Test Performance.
Source: Adapted from Messick (1982, p. 88).

Future Directions

The true measure of a test's reliability and validity is its usefulness in the local context. As language programs select tests for admission, placement, certification, and other functions, it is incumbent on leadership to determine the extent to which the tests are appropriate for and provide useful information to the local program and its participants, from students to instructors to administrators to future employers. The extent to which standardized test results provide useful information can, and should, be demonstrated empirically by program administrators and assessment specialists. Some research has explored student and instructor perceptions of large-scale, standardized tests (Gardiner & Howlett, 2016; Llosa & Malone, 2017), and such research can help institutions make informed decisions about test selection.

New language tests and new testing technology emerge regularly, and these changes to the landscape will always factor into decisions about which tests to use and which scores to accept for different purposes. As discussed previously, the test development process should be meticulous and reflect the authentic language that is part of the target language use domain; in other words, the content of language tests should reflect real-world language production and interaction. While we have referenced the costs of many standardized language tests, we have also tried to underscore the rigor that underlies most standardized tests and the long-term cost savings to students, programs, and institutions by using tests with high reliability, strong validity arguments, and robust research that supports the purpose(s) for which programs may use a test.

Recently, the language testing field has begun to explore expert perceptions about issues such as artificial intelligence in language assessment from automated scoring to item development. The consensus is that, while AI can support and even improve human efforts, humans still must retain a role in high-stakes test scoring (Xi, 2024). Most AI scoring algorithms are developed based on human scoring patterns, and when human judgment is significantly reduced or eliminated from the process, the system not only becomes less reliable but also reflects less human judgments of other humans' language.

A second, and equally important, emerging issue is that of the role of AI in test development. At this writing, AI is still limited in its ability to develop texts and passages at specific proficiency levels with appropriate multiple-choice items. The issue is complex: what is authentic language in a world populated by AI? What message do we send learners when we use or do not use AI? The issue of language variation looms large in both AI scoring and item development. How can we be sure that the corpora used to train AI scoring and to develop includes different varieties of the language and represents authentic human language?

Conclusion

Testing provides opportunities for stakeholders in higher education to make decisions about admissions, placement, achievement, and certification, among other uses. Within language programs, it is critical to ensure that the tests being used are appropriate for the specific purpose and audience of a particular program, that they mirror authentic language use, and that they reflect best practices in test development, administration, and maintenance. Such efforts require more background, time, and attention than many program directors have and certainly would benefit from administrative support to call upon language testing experts to help inform such decisions. This chapter has described general trends in standardized test development as well as two tests commonly used in higher education (TOEFL and ACTFL

OPI). While there are many tests currently available and new exams that continue to emerge, it is critical to ensure that the tests selected for a program reflect the program's audience, goals, and values.

References

Alpine Testing Solutions. (2020). *Examination of the ACTFL Oral Proficiency Interview® (OPI) in Korean, French, and Mandarin for the ACE Review – Part B: Statistical Analysis & Evidence of Validity*. Alpine Testing Solutions.

Bachman, L. F. (1988). Problems in examining the validity of the ACTFL Oral Proficiency Interview. *Studies in Second Language Acquisition*, *10*, 149–164.

Bachman, L., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Barnwell, D. P. (1996). *A history of foreign language testing in the United States: From its beginnings to the present*. Bilingual Press.

Biber, D., & Gray, B. (2013). Discourse characteristics of writing and speaking task types on the TOEFL ibt® test: a lexico-grammatical analysis. *ETS Research Report Series*, *2013*(1), i-128. Gray.

Biber, D., Reppen, R., & Staples, S. (2017). Exploring the relationship between TOEFL iBT scores and disciplinary writing performance. *Tesol Quarterly*, *51*(4), 948–960.

Brown, H. D., & Abeywickrama, P. (2018). *Language assessment: Principles and classroom practices*. Pearson Education.

Carr, N. T. (2011). *Designing and analyzing language tests*. Oxford University Press.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2011). *Building a validity argument for the test of English as a Foreign Language* ™. Routledge.

Cubillos, J. (2010). Computer-mediated oral proficiency assessments: Validity, reliability and washback. *International Journal of Technology, Knowledge and Society*, *6*(6), 85.

Cummins, J. (1979). Cognitive/Academic language proficiency, linguistic interdependence, the optimum age question and some other matters. *Working Papers on Bilingualism*, No. 19.

Dandonoli, P., & Henning, G. (1990). An investigation of the construct validity of the ACTFL proficiency guidelines and oral interview procedure. *Foreign Language Annals, 23*, 11–22.

Davidson, F., & Lynch, B. K. (2008). *Testcraft: A teacher's guide to writing and using language test specifications*. Yale University Press.

Dorans, N. J. (2004). Equating, concordance, and expectation. *Applied Psychological Measurement*, 28(4), 227–246. https://doi.org/10.1177/0146621604265031.

Elliot, M., Blackhurst, A., O'Sullivan, B., Clark, T., Dunlea, J., & Saville, N. (2021). Aligning IELTS and PTE-Academic: A measurement study. In N. Saville, B. O'Sullivan & T. Clark (Eds.), IELTS partnership research papers: Studies in test comparability series (No. 2, pp. 42–64). IELTS Partners: British Council, Cambridge Assessment English and IDP: IELTS Australia.

ETS (2024). TOEFL iBT Test Content (ets.org) accessed July 26, 2024.

Foster, S. K., Paulk, A., & Dastoor, B. R. (1999). Can we really teach test-taking skills? *New Horizons in Adult Education and Human Resource Development, 13*(1), 4–12.

Gaertner, M. N. (2022). *Norm-referenced assessment*. Routledge. https://doi.org/10.4324/9781138609877-REE13-1

Gagen, T., & Faez, F. (2024). The predictive validity of IELTS scores: A meta-analysis. *Higher Education Research & Development*, *43*(4), 873–888.

Gardiner, J., & Howlett, S. (2016). Student perceptions of four university gateway tests. *University of Sydney Papers in TESOL*, *11, 67–96*.

Halleck, G. B. (1996). Interrater reliability of the OPI. Using academic trainee raters. *Foreign Language Annals, 29*, 223–238.

Harsch, C., Ushioda, E., & Ladroue, C. (2016). Investigating the predictive validity of TOEFL iBT® scores and their use in informing policy in a UK university setting. *TOEFL Research Reports*.

Hughes, A. (2003). *Testing for language teachers*. Cambridge University Press.

Ihlenfeldt, S. D., & Rios, J. A. (2023). A meta-analysis on the predictive validity of English language proficiency assessments for college admissions. *Language Testing*, *40*(2), 276–299

International Language Testing Association. (n.d.). *ILTA code of ethics and guidelines for practice.* https://www.iltaonline.com/page/CodeofEthics

Johnson, M. (2001). *The art of nonconversation: A reexamination of the validity of the Oral Proficiency Interview*. Yale University Press.

Knoch, U., & Fan, J. (2024). Test score comparison tables: How well are they serving test users? *Language Testing*, (online first). https://doi.org/10.1177/02655322241239348.

Lantolf, J. P., & Frawley, W. (1988). Proficiency: Understanding the construct. *Studies in Second Language Acquisition, 10*, 181–195.

Liskin-Gasparro, J. (2025). Historical perspectives on the ACTFL Oral Proficiency Interview. In L. Davis & J. M. Norris (Eds.), *Challenges and Innovations in Speaking Assessment* (pagination pending). Routledge.

Liskin-Gasparro, J. E. (2003). The ACTFL proficiency guidelines and the Oral Proficiency Interview: A brief history and analysis of their survival. *Foreign Language Annals, 36*, 483–490.

Llosa, L., & Malone, M. E. (2017). Student and instructor perceptions of writing tasks and performance on TOEFL iBT versus university writing courses. *Assessing Writing*, *34*, 88–99.

Llosa, L., & Malone, M. E. (2019). Comparability of students' writing performance on TOEFL iBT and in required university writing courses. *Language Testing*, *36*(2), 235–263.

Lowe, P. (1988). The unassimilated history. In P. Lowe & C. Stansfield (Eds.), *Second language proficiency assessment: Current issues* (pp. 11–51). Prentice Hall Regents.

Martel, J. (2019). Washback of ACTFL's integrated performance assessment in an intensive summer language program at the tertiary level. *Language Education & Assessment*, *2*(2), 57–69.

Messick, S. (1982). Issues of effectiveness and equity in the coaching controversy: Implications for educational and testing practice. *Educational Psychologist, 17*(2), 67–91. https://doi.org/10.1080/00461528209529246

Schoepp, K. (2018). Predictive validity of the IELTS in an English as a medium of instruction environment. *Higher Education Quarterly*, *72*(4), 271–285.

Shanks, D. R., Don, H. J., Boustani, S., & Yang, C. (2023). Test-enhanced learning. *Oxford Research Encyclopedia of Psychology.* doi: 10.1093/acrefore/9780190236557.013.908

Spolsky, B. (2000). Language testing in The Modern Language Journal. *The Modern Language Journal*, *84*(4), 536–552.

Surface, E.A., & Dierdorff, E.C. (2003). Reliability and the ACTFL oral proficiency interview: Reporting indices of interrater consistency and agreement for 19 languages. *Foreign Language Annals*, 36(4), 507–519.

Thompson, I. (1995). A study of interrater reliability of the ACTFL Oral Proficiency Interview in five European languages: Data from ESL, French, German, Russian, and Spanish. *Foreign Language Annals, 28*, 407–422.

Thompson, I. (1996). Assessing foreign language skills. Data from Russian. *Modern Language Journal, 80*, 47–65.

van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral Proficiency Interviews as conversation. *TESOL Quarterly, 23*, 489–508.

Verhelst, N., Van Avermaet, P., Takala, S., Figueras, N., & North, B. (2009). *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press.

Vyn, R. (2024). Leveraging the AAPPL to promote positive washback in K-12 language teaching. *Foreign Language Annals* (online first). doi: 10.1111/flan.12761

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications.* Cambridge University Press. https://doi.org/10.1017/CBO9780511618765

Xi, X. (2024, July). The voting results are out! [Post]. Linked In (9) Post | Feed | LinkedIn.

Yeom, S., & Llosa, L. (2024). Comparability of reading tasks in high-stakes English proficiency tests and university courses in Korea. *Language Testing in Asia*, *14*(1), 8.