English language tests for university admissions

Test selection guide

Contents

About this guide

This test selection guide is designed to help university administrators evaluate and select English language tests and scores for use in admission decisions. The guide uses principles for test selection developed for the university sector and is based on the International Language Testing Association Guidelines for Practice.

Why evaluate language tests?

Although many tests claim to assess Academic English for university study, test methods vary considerably. Some tests may emphasise word-level tasks such as gap-fill exercises, while others might give more score weight to whether test takers understand complex arguments in longer texts. It is important to choose tests that model academic language as closely as possible and that are trustworthy measures for making decisions about university admissions so that applicants are treated fairly and admissions decisions are based on robust evidence. Ultimately, using more trustworthy test scores for admission decisions gives institutions a level of confidence that commencing students have a level of English that enables them to participate meaningfully in their studies.

Independent evaluation of tests by institutions is necessary so that tests are selected judiciously for typical intake cohorts and particular institutional contexts. The basis for selection should be **detailed information about the test**, **test sample materials** (i.e., what exactly doing the test entails), **research evidence** about the test validity and **technical reporting**. Examining the test itself, rather than relying on the decisions of other institutions or company marketing, is important. Although tests might be accredited by an accrediting agency, these agencies do not typically accredit tests for particular uses in particular institutional contexts.

Most tests used for admissions decisions are commercial products. By nature, all products are subject to sales pressures. While market competition can be healthy, cost-cutting in tests can mean reductions in things like human involvement and test length. Unlike other kinds of products which might have obvious faults that lead to product recalls, safety alerts or reputational damage, language tests that do not work well generally have hidden effects. False positive scores have the potential to admit students who do not have the English skills to cope with English-mediated disciplinary study. False negative scores are unfair to applicants who should have been admitted. For tests that do not simulate the important aspects of the academic domain, applicants may spend years doing "test-wise" preparation activities that are unlike the kinds of language-mediated activities they will need to do at university. This is a waste of time, and frequently of money too.

This guide is a tool to help institutional administrators decide which tests to use for their specific contexts. All major tests claim to be "trusted", but score users (e.g., an institution that uses a particular score threshold for admissions) have a responsibility to determine if a test is "trustworthy" for their institutional contexts.

Structure of this guide

- The **How to use this tool** section outlines an evaluation procedure for selecting English language tests for admissions decisions.
- The **Principles for English Language Test Acceptance** articulate necessary qualities for the valid use of language test scores in university admissions decisions. These can be used as criteria in the test evaluation process.
- The **Test Evaluation Tool** provides explanations and questions based on the principles. The tool lists the information and evidence to gather or request from the **test provider** (i.e., the company or agency that produces the test).

How to use this tool

Evaluating tests requires close engagement with information about the test/s. It is ideally carried out by a committee who are informed about the English language demands of your institutional context. The steps below are suggested as a procedure for determining which tests best match the needs of students with English as an Additional Language (EAL) who are commencing tertiary study in an English-medium context.

1. **Form an evaluation committee.** Various perspectives can be useful, including from people who are (1) knowledgeable about the **demands of academic study via EAL** in your institution, (2) familiar with the institution's **admissions procedures**, (3) familiar with **research and best practice in language testing,** and (4) **are experienced in doing the test/s** and **studying at the institution**. The evaluation committee should not include representatives from test companies.
2. **Gather information and evidence.** Compile relevant information about test/s being considered for the evaluation committee. If information isn't available about an aspect of the test that is of interest to the institution, it is reasonable to request it from the testing company so that the evaluation committee is well informed. "Information and evidence to gather" is listed for each principle in the Test Evaluation Tool. Red flags (⚑) are possible reasons for exclusion based on unavailability of information. There are four main sources, listed below.
    i. **Descriptive information** about the test content from test provider (i.e., testing company) websites, including information about how the test is scored.
    ii. **Sample test materials** are an essential source of information about the actual test content and the experience of doing it. Only official test materials should be consulted. A test provider that does not publish full sample materials for all test sections should not be considered. It should not be necessary to pay or even sign up to a company's database to access official sample materials.
    iii. **Research about each test** should be freely available on the test provider's website. Some of this research should be independent, i.e., carried out by researchers outside the testing company.
    iv. **Quality assurance procedures and technical reports** should be available in the form of descriptions about regular procedures (e.g., rater training methods, test

monitoring methods) and routine reports of test performance in administrations which include measures such as reliability statistics and test taker population investigations.

3. **Evaluate test/s independently.** Using the collected information, committee members independently evaluate each test against the questions in Test Evaluation Tool, and make an overall assessment for each section.
4. **Compile committee evaluations.** Compile the *overall evaluations* from each committee member in an overall evaluations table.
5. **Discuss and select test/s.** Committee members discuss their individual evaluations and seek consensus on the suitability of each test.
   - Principle 1 – *The test is fit for purpose and provides sufficient information for admissions decisions* – is a hurdle principle. If it is not met, the test can be excluded.
   - If there are clear gaps in information available about a test, the committee may decide to request further information or exclude the test. For test content and sample materials, always use official information from test providers.
   - Some aspects of tests and test research are difficult for non-specialists to evaluate. However, the committee can check that the listed components and procedures are publicly available, and in some cases, that transparent procedures are communicated. Committee members may choose not to evaluate parts they feel unable to.
   - Ultimately the decision to accept a test for admissions purposes is based on all principles, on balance.

Principles for English Language Test acceptance

English Language Tests must meet the following criteria before they can be considered for inclusion in an admissions policy.

Principle 1: The test is fit for purpose and provides sufficient information for admissions decisions. (P1)

- The test measures academic English language proficiency: tasks and scoring methods represent similar tasks, skills and values to those of the academic domain.
- The test gathers sufficient evidence about all four skills: reading, writing, listening and speaking.
- Humans are meaningfully involved in evaluating each test taker's use of language.
- The test provider has a validation research agenda which examines the accuracy and appropriateness of scoring, the test's fit-for-purpose for the academic domain, and the impact of the test. The validation research is available.
- The test generates preparation activities that develop academic English.
- The test levels are representative of the range of English language proficiencies relevant to Australian higher education admission.

Principle 2: The test provides consistent and fair measurement. (P2)

- The test is reliable and publishes information about its internal properties.
- The Standard Error of Measurement (SEM) is acceptable at score decision points, i.e., the score/s required to enter programs in the institution.
- Scoring practices are unbiased and fair.
- Appropriate accommodations for test takers with special needs are provided.
- The test is secure.
- Test administration procedures are standardised.

Principle 3: The test's score equivalence with other tests is robust. (P3)

- There is a sound empirical basis for the score equivalences claimed.

Principle 4: Communication and reporting are clear, accessible and timely. (P4)

- Information about the test content and test-taking procedures is publicly available and accessible.
- Sample test materials are freely available and easily accessible.
- Test results are communicated clearly and in a timely manner.
- Appeal processes are clearly described in information for test takers.

Test evaluation tool

Principle 1: The test is fit for purpose and provides sufficient information for admissions decisions

P1.1 **Fit for purpose**	**The test measures academic English language proficiency: tasks and scoring methods represent similar tasks, skills and values to those of the academic domain.** **The test gathers sufficient evidence about all four skills: reading, writing, listening and speaking.** **Humans are meaningfully involved in evaluating each test taker's use of language.**

P1.1 Explanation

Academic English tests should comprise tasks that require test takers to **engage with and produce language in academic genres on pseudo-academic topics**. ⚑A test that does not measure academic English proficiency using substantial, relevant test tasks should not be considered. By nature, tests are time constrained, but academic study requires students to engage with and produce long and complex texts on disciplinary topics. A trustworthy test balances task variety with text and response length. It allows enough time to collect reasonable evidence of a test taker's ability, and that evidence includes a variety of opportunities to demonstrate academic language ability across different types of tasks. A trustworthy test does not waste precious test time with irrelevant tasks/items that might be efficient to score but demonstrate only a narrow aspect of academic language. It is also necessary to look closely at **scoring methods** to check that response complexity is assessed, and not just, for example, number of words or relevance of vocabulary. As with any high-stakes decision-making about people's life opportunities, human judgement should be meaningfully included in the scoring of individual performances, especially for more complex tasks such as academic summaries or arguments.

P1.1 Information & evidence to gather

Test providers usually publish descriptive information and official sample (practice) materials on their websites. This information varies in terms of its quality and detail. If information is not available, it is reasonable to request it, but at a minimum, the following information should be publicly available and gathered for the evaluation.

- Descriptive information about the test content (texts, items, tasks, etc.)
- Full official sample materials for the whole test
- The time allowed for the whole test and sub-tests
- The scoring criteria and how these are operationalised by either human or automated methods
- The types of English represented in the test (e.g., accents)
- Mode/s of delivery (e.g., online, paper, typed, handwritten)

P1.1 Evaluation

Questions	Evaluation of [test name]
GENERAL: To what extent do test takers have to **understand** and **produce** academic English in reading, writing, listening and speaking tasks?	
GENRES & REGISTERS: Is there a variety of **academic genres and texts** (e.g., essay, summary, article, abstract, lecture, presentation, description)? Is there an authentic range of **accents**?	
TOPICS: Is there a range of **academic topics across disciplines**?	
ITEMS: Is there a variety of item types (e.g., multiple choice, short/long answer, matching, cloze).	
TEXT LENGTH: **How substantial** (i.e., how long and complex) are the **written or spoken texts** that test takers have to understand and produce?	
ACADEMIC LANGUAGE ABILITY: Do the tasks require test takers to: • understand and use **academic vocabulary**? • understand written and spoken **academic argumentation and discussion**? • **produce sustained, meaningful speech** connecting their own or others' ideas? • **produce written texts** which require academic skills such as argumentation, use of evidence, summarising or describing?	
SCORING: Do the **scoring methods** value academic language use, for example the coherence of an argument or the range of academic vocabulary used? Is **human judgement** used meaningfully in the scoring process for each test taker, especially regarding complex language use such as academic argumentation and extended discussion?	
P1.1 Overall evaluation: Test content is substantially representative of the academic domain. Delete as applicable: **Agree – Partially agree – Disagree**	

P1.2 Fit for purpose	**The test provider has a validation agenda which examines the accuracy and appropriateness of scoring, the test's fit-for-purpose for the academic domain, and the impact of the test. The validation research is available.** **The test generates preparation activities that develop academic English.**

P1.2 Explanation

It is important that research is conducted to show the extent to which a test is an appropriate basis for decisions about entry into the academic domain. If a test is revised, research should be conducted to show that the test remains relevant.

P1.2 Information & evidence to gather

Research should be available on the test provider's website. ⚑A test that does not have relevant, publicly available validation research should not be considered. ⚑A test that only has company-produced research available should not be considered. Research that relates to superseded forms of the test may not be relevant (e.g., if a test has moved to automated scoring or changed task design since the publication). Research about test validity based on superseded test score mechanisms or task designs should not be included in the evaluation.

Research on the following is useful:

- Test tasks and the skills, knowledge and processes elicited by the test tasks
- Scoring methods, automated/human ratings
- Test preparation activities (called "washback")

P1.2 Evaluation

Questions	Evaluation of [test name]
Does research demonstrate that the current test tasks and topics capture academic language ability? For example, are there studies linking the test tasks to real life academic tasks?	
Is there any research investigating whether the skills, knowledge and processes elicited by the test tasks are similar these are to skills, knowledge and processes of the academic domain (i.e., what the test-takers are thinking and doing while engaging with the test tasks)?	
Does research demonstrate that the scoring places value on academic English ability? For example, is there evidence justifying the scoring criteria used for writing/speaking?	
Does research show that any automated methods used in the test scoring methods are well aligned with human judgements?	

If automated scoring methods are used, is there evidence that the automated methods are strongly representative of the criteria?	
Does research demonstrate that people preparing for the test do activities that are similar to academic tasks? ⚑Is there evidence that the test methods cause people preparing for the test to focus on superficial techniques such as speaking loudly, speaking without pausing, or using as many multisyllabic words as possible?	
P1.2 Overall evaluation: Research evidence indicates that the test is representative of the academic domain. Delete as applicable: **Agree – Partially agree – Disagree**	

P1.3 Fit for purpose	**The test levels are representative of the range of English language proficiencies relevant to the applicant pool.**

P1.3 Explanation

The test should be well suited to the English language proficiency level of a typical commencing cohort of students who will be studying through English as an additional language. The score/s used for admission decisions should not be at extreme ends of a test's scoring scale.

P1.3 Information & evidence to gather

Test provider's information about the scores, including:

- The scoring scale used
- Descriptive information about the language proficiency levels in relation to scores on the scale
- Sample performances at different levels, e.g., a writing sample that is typical of a test taker at different score levels

P1.3 Evaluation

Questions	Evaluation of [test name]
Does the test score range cover the score levels required for admission? ⚑If the score/s used for admissions decisions are at the end of the scoring scale, the test should not be used.	
Do the description and sample performance align with institutional expectations of the proficiency level of commencing students?	
P1.3 Overall evaluation: Test score range and samples are well suited to expected proficiency range of commencing students. Delete as applicable: **Agree – Partially agree – Disagree**	

Principle 2: The test provides consistent and fair measurement.

P2.1 **Consistency & fairness**	**The test is reliable and publishes information about its internal properties.** **The Standard Error of Measurement (SEM) is acceptable at score decision points, i.e., the minimum required score for entry to a program.**

P2.1 Explanation

Consider the overall test reliability and sub-test reliability (e.g., reliability of the speaking, listening, reading and writing sub-tests). For high-stakes decisions, such as university entry, reliability values of 0.8 or higher are considered best practice.

Scores around decision-making points for admission should be statistically accurate. The standard error of measurement (SEM) can be used to establish this accuracy. A smaller SEM indicates a higher degree of accuracy.

P2.1 Information & evidence to gather

Test provider produces regular reports on the performance of different versions and administrations. ⚑A test that does not make reliability statistics for each sub-test and for the overall test score available, should not be considered. For evaluation of Principle 2, seek routine reporting of the following aspects:

- Reliability statistics and internal psychometric properties of test versions.
- Average standard error of measurement (SEM) values at all score levels.
- Monitoring procedures for human and automated rating/scoring methods.

P2.1 Evaluation

Questions	Evaluation of [test name]
Is reliability routinely reported for test versions?	
Are reported reliabilities for the overall test and sub-tests at 0.8 or above?	
Is the standard error of measurement (SEM) relatively small around scores used for admission decisions?	
P2.1 Overall evaluation: Reliability and error measurements are reported and acceptable. Delete as applicable: **Agree – Partially agree – Disagree**	

P2.2 Consistency & fairness	**Scoring practices are unbiased and fair.** **Appropriate accommodations for test takers with special needs are provided.**

P2.2 Explanation

Tests should carry out routine investigations into whether the test is biased towards any particular group. Another safeguard against bias is meaningful human engagement in task development processes, including human review for bias and content that is harmful, discriminatory or derogatory. Tests should have a transparent score appeal procedure which involves a genuine second assessment (e.g., human review of automated scoring, a different human rater for human scoring). Test accommodations should be available so that all test takers can participate equitably and to the best of their ability.

P2.2 Information & evidence to gather

Test provider produces regular summaries of test bias investigations on current versions of the test, e.g., country of origin, gender.

Test provider specify the accommodations available for test takers and communicate these clearly on their test website.

Test provider has transparent procedures for:

- Score resolution processes (in the event of discrepancies between raters or rating mechanisms)
- Appeals processes, including human scrutiny for automatic scoring.

P2.2 Evaluation

Questions	Evaluation of [test name]
Is **test bias** routinely investigated and reported on by the test provider?	
Are **test accommodations** available?	
Do **score appeals** processes involve a genuine second assessment?	
P2.2 Overall evaluation: Procedures for bias detection, test accommodations and score appeals have transparency and integrity. Delete as applicable: **Agree – Partially agree – Disagree**	

P2.3 Consistency & fairness	**The test is secure.** **Test administration procedures are standardised.**

P2.3 Explanation

Scores on a test that is compromised by cheating or lack of security cannot be trusted so it is important that tests have comprehensive security measures. Cheating is possible in all forms of test delivery, but test centre delivery significantly reduces the range of possible methods of test compromise that exist in uncontrolled environments.

Large scale tests cannot produce consistent scores if they are administered inconsistently in test centres across the world.

P2.3 Information & evidence to gather

Test providers publish information about security and integrity procedures, such as:

- Identity checking and test invigilation methods
- Storage of test materials
- Test taker information
- Test data
- Test administration security
- Test administration staff recruitment
- Score report verification processes

Test providers describe how standardised procedures are maintained across test centres.

P2.3 Evaluation

Questions	Evaluation of [test name]
Are there **comprehensive test security** and **integrity procedures** in place for test taker identification, test materials, test data, test taker information, test administration procedures and staff recruitment?	
Are **standardised procedures of test administration** maintained across test centres?	
Are **score report verification** procedures available?	
P2.3 Overall evaluation: The test is secure and has standardized administration procedures. Delete as applicable: **Agree – Partially agree – Disagree**	

Principle 3: The test's score equivalence with other tests is robust.

P3 **Equivalence**	**There is a sound empirical basis for the score equivalences claimed.**

P3 Explanation

Different tests use different methods of eliciting English language ability from test takers. They have different scoring methods, score ranges and score scales. Therefore, scores on tests cannot be compared directly. To determine whether a score on a different test is equivalent, test providers must have empirical evidence from a "concordance study" using a robust methodology and a large sample of test takers around the score level/s required for admissions.

P3 Information & evidence to gather

The test provider publishes one or more test score equivalence studies ("concordance studies") which shows how the test was equated with other test/s, the equivalent scores and any limitations (e.g., small sample size at certain score levels). ⚑A test that does not have a publicly available concordance study available should not be considered.

P3 Evaluation

Questions	Evaluation of [test name]
Does the test have a strong evidence base for its score equivalence to other test/s accepted for admission?	
P3 Overall evaluation: Score equivalences have a sound evidence base. Delete as applicable: **Agree – Partially agree – Disagree**	

Principle 4: Communication and reporting are clear, accessible and timely.

P4 **Communication**	**Information about the test content and test-taking procedures is publicly available and accessible.** **Sample test materials are freely available and easily accessible.** **Test results are communicated clearly and in a timely manner** **Appeal processes are clearly described in information for test takers.**

P4 Explanation

Good communication about the nature of a test and its methods of scoring is important so that test takers know what to expect and can do their best.

Score reports should clearly indicate overall and subtest scores and relevant level descriptions so that the score can be interpreted. The timing of test results depends on how complex the marking/rating process is and the extent to which humans are involved. Test scoring that is fully automated can be reported relatively quickly. Human involvement and checking processes take time. Test scoring (marking/rating/scoring) should have some human involvement, especially for the evaluation of more complex criteria, and there should be checks in the process. Therefore, while immediate scoring is not feasible, results should be returned in a reasonable timeframe (around 4 weeks).

There should be clear information for test takers about what to do if they would like to appeal a score, and the allowed timeframe. The process of review should be transparent.

P4 Information & evidence to gather

Test provider website has the following information freely available and accessible:

- Explanation of the intended use/s of the test score
- Descriptions of all test components (sub-tests, tasks, item types, response expectations, timings)
- Clear, accessible, descriptions of scoring methods (e.g., rubrics, relative weightings of tasks/criteria, explanations of automated methods)
- Score scale and meaning of score levels
- Examples of typical performances at different proficiency levels
- What to expect during the test administration (e.g., identity check, security)
- Accommodations available
- Free official sample materials of full test. ⚑A test that does not provide free, official sample materials should not be considered.

Test provider website provides a sample score report and an indication of expected timeframe for test results.

Test provider website has the following information:

- Appeal procedure for test takers to follow and timing
- Description of score review procedure
- Timeframe for acknowledgement of complaints
- Timeframe for investigation and action plan to resolve complaints

P4 Evaluation

Questions	Evaluation of [test name]
Is information available in accessible language about test structure, test components (sub-tests, tasks, response expectations), cost, scoring criteria and methods, test conditions, score interpretation and meaning, and intended uses of scores?	
Are sample materials of the full test freely and publicly available to test takers?	
Do score reports contain clear, interpretable information about overall scores, subtests and score explanation?	
Do score reports contain information about the test administration, e.g., date, test centre?	
Are results available in a timely manner?	
Is there a clear appeal procedure with timeframe available to test takers?	
P4 Overall evaluation: Information about the test content, test-taking procedures and reporting is clear and accessible. Delete as applicable: **Agree – Partially agree – Disagree**	

Overall evaluations

Compile evaluations (Agree, Partially agree, Disagree) from all committee members as a basis for discussion and consensus.

Overall evaluation	Committee Member 1	Committee Member 2	Committee Member 3	Committee Member 4
P1.1 Overall evaluation: Test content is substantially representative of the academic domain.				
P1.2 Overall evaluation: Research evidence indicates that the test is representative of the academic domain.				
P1.3 Overall evaluation: Test score range and samples are well suited to expected proficiency range of commencing students.				
P2.1 Overall evaluation: Reliability and error measurements are reported and acceptable.				
P2.2 Overall evaluation: Procedures for bias detection, test accommodations and score appeals have transparency and integrity.				
P2.3 Overall evaluation: The test is secure and has standardized administration procedures.				
P3 Overall evaluation: Score equivalences have a sound evidence base.				
P4 Overall evaluation: Information about the test content, test-taking procedures and reporting is clear and accessible.				
Tentative consensus (for discussion)				

For questions or further information, please contact the authors:

Ute Knoch, Language Testing Research Centre, University of Melbourne
uknoch@unimelb.edu.au
Susy Macqueen, Australian National University susy.macqueen@anu.edu.au

21 August 2025