BRIEFING SHEET

www.ukalta.org info@ukalta.org @UKALTA2

No. 1/23, July 2023

WHAT IS LANGUAGE TESTING?

This briefing sheet answers the following questions:

- What is language testing?
- What are some key concepts in language testing?
- What are some new trends in language testing?

AUTHOR

Nathaniel Owen, Oxford University Press, Oxford.

WHAT IS LANGUAGE TESTING?

Language testing is the practice of evaluating language for the purposes of certification or decision-making.

Language assessment refers to any activity in which information is collected from language learners from which we make judgements about their language proficiency or progress.

Assessment can therefore encompass both informal, classroom-based quizzes, self-assessment or peer assessment and more formal, high-stakes summative assessment. Testing is typically the more specific term for formal, standardized assessment. Language tests are usually constructed and provided by examination agencies who are granted the authority to issue certificates which are recognised for specific purposes such as entry to higher education, or by employers who wish to see evidence of proficiency as part of decision-making for prospective employment.

WHAT ARE SOME KEY CONCEPTS IN LANGUAGE TESTING?

Some key concepts in language testing include (but are not limited to) **test purpose**, **stakes**, **validity**, **reliability** and **fairness**.

- **Test Purpose.** Any individual test should always have a purpose which can be clearly explained by the test creators. For example, a low-stakes classroom test could be *diagnostic*; that is, created and implemented to identify areas requiring further learning or remedial action. Alternatively, classroom-based tests might be evaluative, to make claims about proficiency at the end of a course. Additionally, this test might be used to make decisions based on a *cut score*, such as for admission to higher education. Other purposes for testing might be to place learners in suitable classes (a placement test) or a progress test based on a subset of course learning objectives. Each test is made up of *prompts*, which elicit a response from a test taker. This response is the *evidence* that we use to inform decision-making. This evidence may be used *formatively*, to enable learners to judge their progress. In proficiency tests, the evidence may be used *summatively*, which means decisions are made about an individual on the basis of performance at the end of a course of instruction.

- **Stakes.** Stakes refers to the *consequences* of test outcomes. Stakes may vary depending on test purpose. For example, classroom-based tests may be *low-stakes,* as the decisions made on the basis of the scores are unlikely to have long-term ramifications for individual test takers. High-stakes tests are those which are used for significant decision-making, such as university admissions, graduation, moving to another country or applying for a job. The stakes of a test will have a significant impact on test takers' motivation, what and how test takers study and even their mental health. Tests should therefore be designed with these effects in mind: to accentuate the quality of the information provided to decision-makers and trying to avert potential negative effects on test takers.

- **Validity.** In testing, validity refers to the extent to which *inferences* about test scores are justified. A validity *argument* sets out the evidence and theory to support claims about test score meaning. The kinds of evidence might include a comparison of the test content with the kinds of content we would find in the 'real world'. Evidence might also include studies relating test scores to external criteria, for example academic performance in the first year of university study. This is known as *predictive validity*. Studies may focus on conversation analysis which shows that a certain test reflects a predicted range of language functions or conversational features. The type of evidence required depends upon the claims made for the scores. The *theory* in the validity argument provides the rationale for claiming that the evidence presented supports the claims made about the test scores. In low-stakes tests, particularly classroom-based assessments, validity evidence may be collected informally. Validity questions we might ask would include: Has the feedback resulted in learner improvement? Are the tasks engaging and challenging for learners at this level? Is learner motivation improving?

- **Reliability.** Reliability refers to the consistency of measurement across facets of the testing context. These facets might include *time, place, interlocutor,* and *rater/marker*. Test scores should not vary depending upon where the test is taken, who the interlocutor may be, or who rates the performance, as these facets are *irrelevant* to what we are trying to measure. We can investigate the impact of these facets on test scores by changing one while holding

others constant and evaluating the impact on test scores. If a test successfully assesses language proficiency, we would not expect test scores to change if the test is taken two or three times over one week by the same test taker (excepting any normal random error variance due to chance factors). However, over longer periods of time, scores may change due to study or attrition. Reliability is therefore related to validity. A test must be reliable in order to be valid. However, reliability is not sufficient evidence on its own to claim validity (after all, it is possible to reliably measure the wrong thing). Validity and reliability are important because they speak to our concern for *test fairness* in all aspects of assessment practice.

- **Fairness.** Fairness in testing means that all learners should have an equal opportunity at the point of assessment to show their proficiency. This means that there should be no *bias* towards or against any subgroups of the population, and that the scores should be *independent* of test method facets that are irrelevant to what we are assessing. Evidence of differential performance among subgroups may lead to charges of discrimination. Consistent with the distinction between reliability and validity, test bias is something that we can investigate statistically. However, demonstrating an absence of test bias is insufficient to claim that the test is *fair*.

Test fairness is an argument made about the defensibility of test use for a specific purpose by stakeholders.

WHAT ARE SOME NEW TRENDS IN LANGUAGE TESTING?

Trends in language testing often mirror those in Applied Linguistics, or in the Education field more generally. Therefore, some recent concerns in language testing include (but are not limited to) multilingualism, on-demand testing and artificial intelligence. (See other UKALTA briefing sheets on some of these topics.)

- **Multilingualism.** Increasing attention is being paid to educational contexts in which learners speak more than one language and draw upon these for their learning experiences. There is concern that the linguistic diversity of these contexts is not well reflected in existing tasks or test content. How should stakeholders and language test developers respond? What kinds of 'Englishes' should be included

in high-stakes tests, and how well are students served by English language tests in these contexts? Additionally, there are questions about the appropriacy of English language tests designed for use in contexts such as the UK, USA or Australia being used in English-medium contexts such as Sweden, Nepal or India, for example. These are countries in which English serves as the language of education, but is not the first language of the students in that context.

- **On-demand testing.** The Covid-19 pandemic resulted in an increase in 'on-demand' or 'at home' testing. At-home tests are much more flexible and convenient for test takers who can take a language test at a time and place of their choosing. However, there are legitimate security concerns associated with the use of at-home testing versus traditional testing with the use of invigilators. At-home tests typically employ online invigilators known as 'remote proctors' to oversee high-stakes language tests. Despite this, there are understandable concerns that remote proctoring may be an excessive infringement on personal privacy if security includes 'room sweeps' by remote proctors using the test takers' own webcam. Remote proctoring is usually supplemented by artificial intelligence solutions, such as test taker eye movements.

- **Artificial intelligence.** Artificial intelligence is swiftly becoming the major focus of research in language testing. Whether for remote proctoring, adaptive testing, or automated scoring, AI-related research is rapidly influencing how we assess language. Discussions around automated scoring have been ongoing for more than a decade, with incremental improvements in score reliability evident. However, there are continuing discussions as to whether the level of agreement between human assessors and machine scoring is sufficient to claim that automated scoring models are reliable. Additionally, traditional models of validity may be too impoverished to encompass the use of sophisticated engines such as Google's BERT or OpenAI's GPT-3. The field is actively looking towards how validity in language testing can inform and be informed by approaches to machine learning. In early 2023, we saw significant discussion about the impact of ChatGPT on Educational Assessment and in Education more generally. In future it is likely that the impact of models such as OpenAI's GPT4.0 or Google's

BERT will significantly influence the direction of language testing and assessment.

Recommended reading:

Douglas, D. (2010). *Understanding Language Testing*. London: Hodder Education/Routledge.

An introductory text that explains basic terminology and key concepts in language testing, outlines the skills required to design and use language tests, and introduces simple statistical tools for test analysis. No prior knowledge of language testing is assumed.

Fulcher, G. (2010). *Practical Language Testing*. Hodder Education/Routledge.

An intermediate text dealing with the purpose of testing in context and an analysis of test use in society. The text then follows the 'test development cycle' to explain in detail the process of test design, implementation, and interpretation.

Carr, N. (2011). *Designing and Analyzing Language Tests*. Oxford University Press.

Provides a comprehensive overview of concepts, principles, and methods for designing, developing, and evaluating language tests. The book aims to equip readers with the knowledge and skills to critically analyse existing tests and thoughtfully develop new language tests that are valid, reliable, practical, authentic and beneficial for test takers.

Green, A. (2020). *Exploring Language Assessment and Testing* (2nd ed.). Oxford University Press.

Covers a wide range of topics related to language assessment, including theoretical foundations, assessment design and development, test administration and scoring, and the use of assessment results**.**

Hughes, A., & Hughes, J. (2020). *Testing for Language Teachers*. Cambridge University Press.

Covers topics such as the purposes of testing, test content, test methods, scoring, analysing test performance, developing tests, ethics, and monitoring standards. Guidance is provided on developing tests, evaluating existing tests, giving feedback on tests, and using tests as part of the teaching process