A Guide to Submitting Validity Evidence

> *6.2.3.3 The overriding concern of high-stakes test developers should be fairness. In language testing, fairness is interpreted in terms of validity and reliability. Practicality is a third fundamental test consideration. All tests should be evaluated in terms of their validity, reliability, and practicality based on documented evidence* (ICAO Doc 9835).

The purpose of this document:

1. To ensure that concise and appropriate evidence in support of the validity argument are submitted to AELTS.
2. To serve as a screening tool for TSPs who are considering submitting their test for evaluation.
3. To serve as an educational tool regarding the general form and purpose of validity evidence required for the AELTS process.

What is "validity"?

Broadly defined, "validity" is the extent to which a test measures what it is intended to measure. More specifically, validity is the extent to which the inferences that testers make based on test scores are justified. In the case of tests designed in response to the ICAO LPRs, the validity argument should ultimately demonstrate that the test tasks and the assessment criteria are aligned with the goal of the ICAO LPRs. In other words, the inferences drawn based on test scores derived from tests developed in response to the ICAO LPRs should be reflective of candidates' ability to communicate in non-routine aviation situations. Ultimately, *"(t)he purpose of the ICAO language proficiency requirements is to ensure that the language proficiency of pilots and air traffic controllers is sufficient to reduce miscommunication as much as possible and to allow pilots and controllers to recognize and solve potential miscommunication when it does occur"* (ICAO Doc 9835, 2010, paragraph 4.2.1).

What is "reliability"?

Reliability refers to the consistency of measurement. Reliability may also be conceived of as an aspect of validity (scoring validity), as any inconsistencies in measurement would constitute a threat to the validity of inferences drawn based on the scores obtained.

How do we know if inferences based on test scores are valid?

Validity arguments must be constructed *based on evidence*. Tests must undergo an evidence-based validation process to ensure that inferences drawn based on test scores are indeed valid.

What is a validation process?

Validity is a multifaceted concept, and different types of evidence are required to support claims made regarding the validity of test scores. Both quantitative and qualitative data and research methods can be used in the validation process. All evidence should be methodically collected, analyzed and reported. Some aspects of the validation process occur before the test event (i.e., in the design and development phase) and other aspects of the validation process occur after the test event (i.e., based on data obtained in the trialing and live testing phases). Validation should be considered an *ongoing process*. For example, testers are often required to return to the design and development phase based on the results of ongoing validation studies conducted after trailing and live testing.

How do I conduct a validation study?

Validation should be a consideration from the initial stages of test development. Ideally, test service providers should involve test development and validation experts ("language testing experts") from the conceptual phase. Once the test has been developed and trialing is complete and/or the test has gone live, an external validation study conducted by reputable language testing professionals can augment the credibility of the validity argument. Some test service providers will submit a report of a full external validation study conducted by reputable professionals. Others have the in-house expertise to conduct their own validation studies and submit reports of those studies. With either approach, concise yet comprehensive reports of each step of the validation process should be submitted in the application to AELTS. Such reports should reflect the methodological approach taken in each validation phase. Methods for data collection and analysis should be reported in addition to the results and the implications for test development, maintenance and/or validation.

What types of validity evidence should I submit?

The validation process may differ depending on the test, so it is not possible to provide an exhaustive list of requirements or a step-by-step guide regarding evidence to be submitted. Nevertheless, below is a list of considerations for test service providers who wish to submit their test to AELTS. Consideration of the items in this list may help test service providers decide whether or not their test is ready for submission to AELTS. It is important to provide a clear rationale and/or empirical evidence in support of all validity claims.

Consider the following issues in your submission:

Test Design and Specifications

1. Have detailed test specifications been provided? Please explain the process of their creation.

2. How often are the test specifications reviewed and updated?

3. How are the criteria of the rating scales and the holistic descriptors (ICAO Doc 9835, and Attachment A to Annex 1) reflected in the format and structure of the test? Explain why this approach was adopted.

4. What process was followed to ensure that the test assesses language specific to the language needs of air traffic controllers or pilots?

5. Describe the rationale for the task types, items and presentation of content in the test instrument. What measures were taken to ensure the tasks **and** type of language elicited in each part/section of the test represents the kind of communication skills/language pilots and/or air traffic controllers actually use in their jobs?

6. If the same test instrument is used to assess both pilots and controllers, how does the test address the specific language skills associated with each profession?

7. What is the relationship between the separate sections/components of the test and the overall score awarded for each of the criteria? Describe the rationale for any segregation of test components. How are they scored? What contribution do they make to the overall scores? Explain the process used to justify this approach.

8. What was done to ensure that the scores awarded for each component of the test are meaningful and represent the most accurate reflection of the skills (criteria) they aim to assess? Describe what methods were used in the design of the test instrument to minimize how scores could be affected by abilities which are not related to the skills it is designed to assess.

9. If the purpose of any component of the test is to evaluate specific criteria rather than all six criteria, what measures were put in place to ensure that the scores only reflect an evaluation of those criteria?

10. If test tasks or items are scored using a system other than the ICAO rating scales, how are these scores converted to or how do they contribute to the final ICAO level awarded? Provide justifications for all decisions made.

11. How is Comprehension assessed? Explain the rationale for the approach adopted.

Trialing and Maintenance

12. When multiple versions of the test were first developed, what checks were performed to ensure they would perform comparably? What was done to verify that each version contained components that assessed the same skills, produced similar overall results and were equivalent in complexity and level of difficulty?

13. If the test contains components/tasks/items which are intended to target language ability at a specific level, what has been done to determine that these components are able to effectively assess this level and discern between other levels?

14. Please describe the trialling process. Have you trialled all items / tasks / versions on a representative sample of test takers?

15. What has been done to establish and confirm the level of difficulty of specific items/tasks/components?

16. How is the number of items/versions required for the target population determined?

17. If the same test instrument is used to assess both pilots and controllers, what measures have been taken to ensure and confirm that the test does not bias in favour of or against a particular group of test takers (e.g. a commercial pilot versus a private pilot versus an air traffic controller)?

18. What changes were made based on the trialling results?

19. Describe the process for the development of new items/tasks.

20. How are item writers trained?

21. When new versions/components/tasks or items are added to the test bank, what process is used to determine if they are suitable for inclusion?

22. What process is adopted to determine if and when components or versions within the test bank need to be updated or replaced? What kind of monitoring is conducted to check that all components of the test bank contribute in the way they were intended and that the scores they generate are consistent with what is expected?

23. How is test item / task performance monitored? What analyses are carried out?

Reliability

24. What checks are used to determine that items/tasks/components of each part of the test instrument perform consistently in the scores they generate?

25. What procedures are used to monitor intra and inter-rater reliability? Explain the basis for any decisions that are made on the outcome of the results of this monitoring.

26. If you use interlocutors during your test, how do you monitor their behaviour and performance?

Some useful resources *

Alderson, C. J., Clapham, C. & Wall, D. (1995). *Language Test Construction and Evaluation* Cambridge*;* Cambridge University Press,.

Bachman, L. F. (2004). *Statistical Analyses for Language Assessment.* Cambridge: Cambridge University Press.

Fulcher, G. & Davidson, F. (2007). *Language Testing and Assessment: An Advanced Resource Book.* New York: Routledge Applied Linguistics.

Lazaraton, A. (2002). *A Qualitative Approach to the Validation of Oral Language Tests.* Cambridge: Cambridge University Press.

McNamara, T. F. (1996). *Measuring Second Language Performance.* London: Longman.

Weir, C. (2005). *Language Testing and Validation: An Evidence-Based Approach*. Palgrave-New York: Macmillan.

* This is not an exhaustive list.