BRIEFING SHEET

No. 2/22, November 2022

WRITING ITEMS FOR LANGUAGE TESTS

This briefing sheet answers the following questions:

- ▪ What is item writing and why is it important?
- ▪ How are language tests written?
- ▪ Who can be an item writer?
- ▪ How are item writers trained?

AUTHOR

Olena Rossi (awarded a PhD in Language Testing at Lancaster University)

UKALTA: THE FACE AND HEART OF LANGUAGE TESTING IN THE UK

WHAT IS ITEM WRITING?

Item writing is usually defined as *producing test items (tasks, questions) against a set of specifications* (Davies et al., 1999).

Test specifications are the core document on which item writing for a language test should be based. They outline the purpose and characteristics of each type of item, task or question in the test, and specifications are needed whether the test is a large-scale commercial language test or a more locally produced institutional test for a university or school. Examples of publicly available language test specifications include Specifications for the Hungarian School-leaving Examination in English and Aptis General Technical Manual.

In addition to the test specifications, item writers are sometimes provided with item writer guidelines. These include a level of practical detail that is essential for writing particular item types, but which may not be needed by other test stakeholders.

Item writing is of critical importance in language testing because the quality of test items directly affects test validity.

WHY IS ITEM WRITING IMPORTANT?

Item writing is of critical importance in language testing because the quality of test items directly affects test validity. A test is sometimes said to be valid "if it measures accurately what it is intended to measure" (Hughes, 2003, p.26). What we intend to measure in a test is often called the **test construct**, e.g., a language skill such as reading or speaking ability.

High-quality test items contribute to overall test validity, while low-quality items risk being a validity threat. Two main threats to test validity that are directly relevant to item writing are:

- **construct underrepresentation** – this occurs when an important aspect of the intended construct is not targeted or covered by the test items
- **construct-irrelevant variance** – this occurs when the test scores reflect not only test-takers' knowledge of the intended construct, but also of some other variable(s) that are unrelated to the construct of assessment, e.g., general knowledge.

HOW ARE LANGUAGE TESTS WRITTEN?

High-quality language tests are produced by following a careful test design process (or cycle) such as the one shown in the graphic below. The process starts from a clear understanding about the intended purpose of the test.

Test design cycle. Adapted from Fulcher, 2010, p.94

In the graphic above, 'writing items and tasks' is represented as a single discrete step. However, item writing is actually an iterative process that follows its own mini-cycle. After the first version of a test item has been produced, it goes through a cycle of **item review**, **revision**, **trialling** (sometimes also called 'piloting' or 'pre-testing') and, if necessary, **further revision** until it is considered ready for inclusion into a live (operational) test.

High-quality language tests are produced by following a careful test design process (or cycle).

During **item review**, draft item and tasks are normally scrutinized by experts (e.g., senior item writers) or item-writing colleagues to check how well the item matches the original test specifications and to determine whether it might perform well in a test. More than one reviewer is usually involved. Reviewers can work independently of each other or as a panel to discuss items together. They often use quality assurance (also sometimes called 'item moderation') checklists to make sure that key aspects of the item or task, such as length, suitability of topic, lexical level and grammatical complexity, have been carefully considered. The Association of Language Testers in Europe (ALTE) provides examples of generic item analysis

checklists for reading, writing, listening and speaking tasks. Because each item type comes with its own set of item writing challenges, it is also useful to have item-type specific checklists, for example for multiple-choice, multiple matching, true-false or short-answer question tasks.

Following this review stage, some items might be rejected, some will be returned to the item writer for **revision**, and some might be considered good enough to proceed to the trialling stage. If an item or task is returned for revision, item writers need an explanation of why the item was problematic and how it can be improved. Revised items can then undergo a fresh round of review by senior experts or a peer panel.

Items and tasks that have passed the review stage can go for **trialling** on a sample of test-takers who should be as representative of the target test population as possible. Test-taker responses from the trialling stage can then be examined to check other aspects of item quality. For example, responses on selected-response tasks (e.g., multiple choice questions) can be analysed statistically to determine whether the level of item difficulty is appropriate and how well the item discriminates between stronger and weaker test-takers. Responses to speaking and writing prompts can be analysed qualitatively to check whether the instructions were clear to test-takers and whether the tasks elicited enough language of the right sort for scoring purposes.

After trialling, a final decision can be made on whether an item is ready for inclusion in a live test. However, it is a good idea to keep monitoring item performance during live testing (see ALTE Quality Assurance Checklists for test analysis and post-examination review). This is because not all item faults can be uncovered during trialling on a limited number of test-takers. Items that have failed to perform well in a live test are often retired.

To sum up, item development is a complex and incremental process that best occurs in a series of iterative steps. Item writers are at the core of the whole process by choosing or creating suitable texts, writing and then revising their items. Item writers might also combine several roles, acting both as writers of their own items and reviewers of their colleagues' items.

WHO CAN BE AN ITEM WRITER?

In large-scale language testing, item writers may be trained language assessment specialists who are employed by a language testing organisation. Some organisations have an item writer recruitment process

and use experienced language education professionals who work as freelancers and are regularly commissioned to produce test items. They sometimes work together in teams on a specific test or skill under the guidance of a more experienced team leader. In university and classroom settings, lecturers and teachers are often required to produce institutional and classroom tests for their students, but usually without any formal training in item writing. Even if trialling is not possible for smaller-scale institutional or classroom-based testing contexts, item writers can still produce good quality items, tasks and questions by consulting some of the advice and guidelines that are now freely available.

The ILTA Guidelines for Practice emphasize the need for item writers to be familiar with current language testing theory and practice, while the EALTA Guidelines for Good Practice in Language Testing and Assessment include considerations of relevant teaching experience and training. In practice, individual organisations' requirements for hiring item writers may vary and in some cases item writers may need a degree, a teaching qualification and several years' teaching experience.

Item writers for language tests do not need to be native speakers of the language but they are normally highly proficient users with some formal knowledge of the language. Good general writing ability is helpful in item writing and successful item writers are aware of techniques to produce particular types of items. Sometimes an item writer may be proficient at writing one type of item (e.g., to test vocabulary) but is less skilled at another type (e.g., a reading comprehension task). Item-writer specialisation can help to capitalise on the relative strengths of each individual item writer.

WHAT TRAINING DO ITEM WRITERS NEED?

To produce high quality items, it is best if item writers are trained. Untrained novice item writers tend to produce poor-quality, flawed, idiosyncratic items, so training item writers contributes to test quality and constitutes important evidence for overall test validation.

In large-scale testing, different assessment organisations take different approaches to item writer training. A generic training weekend might serve as an induction training for item writers and could involve: (1) an overview of the organisation's examinations and an introduction to the principles of test design and production as well as the basic terminology used to describe test questions; (2) two-hour sessions on the techniques of writing particular item types, including input from the trainer and group activities drawing on the

ideas and experience of the participants; (3) an overview of writing for particular skills papers where participants are introduced to how writing for each of these skills impacts on the item type and any implications for the item writer; and (4) techniques for text selection and adaptation. An alternative approach might involve a one-day face-to-face workshop covering: introduction to the CEFR and practice with scale descriptors; selecting texts; technical item-writing principles; sensitivity issues; working with item templates; item reviewing; feedback on acceptance rate and reasons for rejection. Many language testing organisations deliver their item-writing training online. For the British Council Aptis test, novice item writers now train online over a course of several weeks instead of attending a week of face-to-face training.

To produce high-quality items, it is best if item-writers are trained.

Ideally, item-writing training should include both theoretical input into the language assessment principles underlying item writing and practical techniques for producing items for a particular test. But there are also lots of practical resources available nowadays, including freely online, to help anyone wishing to understand the principles and practice of language test design with a view to developing their item writing skills (Fulcher 2010, Hughes and Hughes 2020, Green 2021).

SOME USEFUL REFERENCE RESOURCES

Fulcher, G. (2010). *Practical language testing.* Hodder Education.

Fulcher, G. and Harding, L. (2022). *The Routledge handbook of language testing* (2nd edition). Routledge. (esp, Sections 4-5-6)

Green, A. (2021). *Exploring language testing and assessment* (2nd edition). Routledge.

Hughes, A. and Hughes, J. (2020) *Testing for language teachers* (3rd edition). Cambridge University Press.

Winke, P. and Brunfaut, T. (2021) *The Routledge handbook of second language acquisition and language testing.* Routledge. (esp Parts II and IV)