CAL CENTER FOR APPLIED LINGUISTICS

CAL Commentary
September 2018

Ask a Test Developer

Mina Niu, Kristine Nugent

When people take high-stakes assessments, the results can be used to make important decisions related to their education and career (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). For that reason, it is crucial that assessments are valid, reliable, and fair measures of test takers' abilities. At the Center for Applied Linguistics (CAL), we are committed to developing high-quality language assessments that meet these criteria. This commitment means that we undertake a rigorous test development process for all of the assessments we produce.

Because the results of high-stakes assessments can be used to make important decisions affecting test takers, it is critical that assessments are valid, reliable, and fair.

As test developers, we often get asked about what we do and how we do it. Here are the answers to some frequently asked questions about the test development process at CAL.

Q: Who is involved in the test development process?

A: Assessments developed at CAL begin with the CAL Validation Framework. This framework joins the Assessment Use Argument (Bachman & Palmer, 2010) with Evidence Centered Design (Mislevy, Steinberg, & Almond, 2003; Mislevy, Almond, & Lucas, 2003), which is a comprehensive effort to promote validity throughout the life of an assessment, from initial design considerations through the decisions made on the basis of the assessment and the consequences of the decisions. See Kelly, Norton and Renn (2018) for more information on the CAL Validation Framework and its implications for test developers.

The test development team at CAL works with a variety of individuals to ensure that test development processes embody the CAL Validation Framework. The test design and test development phases are greatly informed by input from educators who have expertise in teaching the examinees who would take the test. This allows us to ensure that the development of item specifications align with the target language use domain experienced by the examinees in the relevant language-learning setting. For example, for the WIDA ACCESS for ELLs 2.0 test, classroom teachers generate themes in the very early stage of item development. CAL refines the initial items and then they are reviewed by teachers.

When the test content is more fully developed, a group of educators reviews the test materials for appropriateness of the content to ensure that no drift has occurred during development, and a different group of educators reviews the same materials for bias and sensitivity issues. As needed, educators take part in piloting, or trying out, new items with students in the target test-taker population.

As you can see, the test development process at CAL requires the collaboration of several key groups. We work with item writers to create the test content, refine items with feedback from educators, and work with a team of artists and programmers to create graphics and layout for the test items. We also develop rater training materials to be used during test administration by trained raters. Finally, when we receive test results, CAL's psychometricians analyze the quantitative data.

Q: How long does the process of creating a test take?

A: The total time varies depending on the test, but generally the process of designing and developing a test can take months or even years. The total time to develop a test depends on a number of factors. For example, the time it takes to define the underlying construct to be assessed and then to create and refine an appropriate test design for the intended use of the test varies greatly. Once the test design and test specifications have been developed, a pool of items could be in the development stage for close to one year, then spend several months as a field test item before being selected to appear on an operational test. For tests with shorter development periods, at a minimum, there are several months of development and review before test takers see a field test form. Figure 1 shows a broad overview of what a test development process could look like. Following a process like this is rigorous and ensures that tests will be of the highest possible quality.

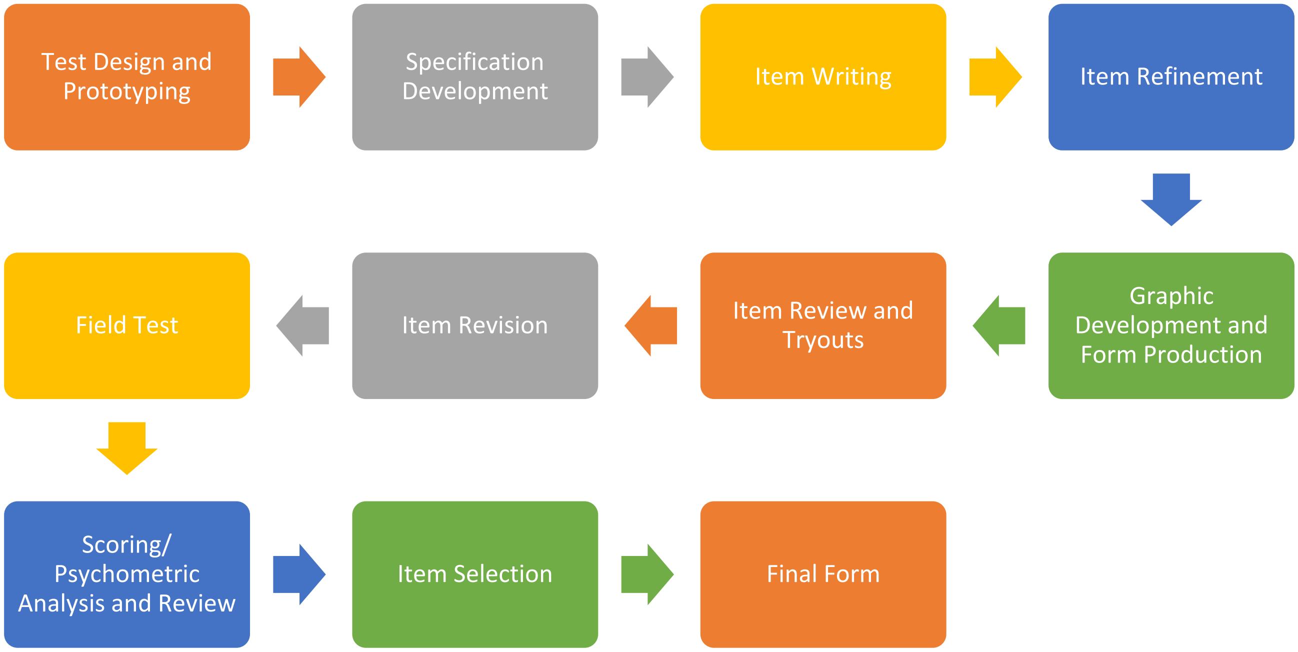
Test Design and Prototyping
Specification Development
Item Writing
Item Refinement
Graphic Development and Form Production
Item Review and Tryouts
Item Revision
Field Test
Scoring/ Psychometric Analysis and Review
Item Selection
Final Form

Figure 1. Test development overview.

Q: How do you ensure that test items are fair?

A: Our goal is that test items are equally accessible to test takers of any background, and that they assess language proficiency fairly. CAL staff are trained to avoid and detect bias when developing test items, and there are stages in the review process during which items are specifically reviewed for bias and sensitivity issues. This includes assembling a panel of educators to review items and identify any potential sources of bias. Additionally, once we have quantitative data from field testing, our statisticians are also able to examine items for potential bias by analyzing how different subgroups perform on each test item. You can read more about CAL's approach to addressing bias here.

Q: How do test developers use field test data?

A: Student response data from the large-scale field tests undergo quantitative analysis. The psychometrics and quantitative research team conducts quantitative analysis to compute statistics which express item difficulty, item fit (i.e., the extent to which test-takers perform as expected, based on their known ability level), item discrimination (i.e., an item's ability to differentiate among students of different proficiency levels), and score distributions for all item types. Test developers use these results to decide which items can appear on an operational test.

The test development team also conducts qualitative analyses of responses to performance tasks (like speaking and writing tasks) to ensure that the tasks are eliciting the type of language that we intended. Only items that have evidence that they are appropriate for operational testing are chosen to appear on final test forms.

Q: What are some special considerations to have in crafting a language proficiency test?

A: It is a challenge to assess language without assessing content knowledge, but our rigorous test development process helps us do this effectively. As prescribed by the CAL Validation Framework, our language tests use authentic, relevant situations as the vehicle to assess language proficiency. For example, test takers might be asked to read a workplace safety poster when taking the BEST Literacy, a literacy test designed for adult English Language Learners (ELLs). However, test takers of WIDA ACCESS 2.0, a test for K-12 ELL students, might read about a science experiment in a school context. Language test items may be contextualized in relevant situations, but they focus on assessing the language related the given situation, not an examinee's knowledge of the content itself.

We keep in mind that test takers come from diverse cultural backgrounds, geographic locations, and socio-economic levels. For example, first graders in Florida and Alaska have very different experiences with weather, so we would not ask them to answer a question that assumes they have personal experience with snow. For adult test takers, this may be a less obvious issue, but we still ensure that on our tests for adults, no outside knowledge of a particular topic is required in order to demonstrate language proficiency.

Q: Why do some tests get new items every year?

A: Test items are refreshed to prevent over-exposure of the items by test takers. This helps ensure test security and validity. Refreshment is also a great opportunity to make sure that test content is up-to-date and aligns well with relevant content standards. In addition, new test items allow us to pilot innovative item types and topics.

The refreshment cycle depends on many factors, including the purpose of the test, the scale of the test, the design of the test, and the test taker population. The test development team at CAL takes these factors into consideration and make new items for tests that require annual refreshment.

Q: What is a typical day like for a test developer?

A: Depending on the stage of the test development process (see Figure 1), a given day could include a wide variety of activities, for example: revising items independently and in teams, meeting with artists to discuss graphics, analyzing field test data, holding online or in-person reviews with educators, and discussing test results with psychometricians. Before new tests become operational, we spend a lot of time checking test content in both paper and online formats to make sure there are no errors and that the online test platform is working correctly. Sometimes we visit schools to try out items with students. Typically, test developers at CAL are working on a number of tasks at once, for tests that are at different points within the test development process.

Takeaways

There is so much that goes into creating valid, reliable, and fair tests. As a reminder, key questions to consider when developing language proficiency tests include:

- What are the purposes of the assessment?
- Who are the test takers and other stakeholders involved in the test?
- Do test items assess language proficiency and not content knowledge?
- Are test items representative of current standards?
- Are test items free of bias and sensitivity concerns?
- How are stakeholders involved in the test development process?

These considerations inform CAL's test development process. Educators and practitioners can also keep these considerations in mind when designing and selecting assessments in order to achieve the best outcome.

References

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

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014*). The Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

Kelly, J., Norton, J., & Renn, J. (2018). Addressing consequences and validity during test design and development: Implementing the CAL Validation Framework. In Davis, J. M., Norris, J. M., Malone, M. E., & Son, Y. A. (Eds.). *Useful Assessment and Evaluation in Language Education* (pp. 185-200). Washington, DC: Georgetown University Press.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, *2003*(1), i-29.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary research and perspectives*, *1*(1), 3-62.

About the Authors

Mina Niu is a Language Testing Specialist at the Center for Applied Linguistics. Before joining CAL, she worked at the Assessment and Evaluation Language Resource Center at Georgetown University. Mina received an M.S. in Applied Linguistics from Georgetown University and a B.A. in Chinese and Bilingual Studies from Hong Kong Polytechnic University.

Kristine Nugent is a Language Testing Specialist at the Center for Applied Linguistics with experience in language teaching and assessment for both adults and K-12 learners. She holds an M.S. in Applied Linguistics from Georgetown University and a B.A. in Romance Languages from the University of Notre Dame.

About CAL

The Center for Applied Linguistics (CAL) is a non-profit organization founded in 1959. Headquartered in Washington DC, CAL has earned an international reputation for its contributions to the fields of bilingual and dual language education, English as a second language, world languages education, language policy, assessment, immigrant and refugee integration, literacy, dialect studies, and the education of linguistically and culturally diverse adults and children. The mission of the Center for Applied Linguistics (CAL) is to promote language learning and cultural understanding by serving as a trusted resource for research, services, and policy analysis. Through its work, CAL seeks solutions to issues involving language and culture as they relate to access and equity in education and society around the globe.

CAL
CENTER FOR APPLIED LINGUISTICS
www.cal.org

Promoting Access, Equity, and Mutual Understanding
Among Linguistically and Culturally Diverse People Around the World