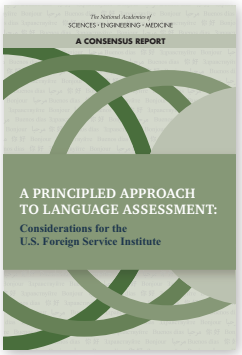
Consensus Study Report

March 2020

HIGHLIGHTS

A Principled Approach to Language Assessment: Considerations for the U.S. Foreign Service Institute



The U.S. State Department needs Foreign Service officers who are proficient in the local languages of the countries where its embassies are located. To ensure that the department's workforce has the requisite level of language proficiency, its Foreign Service Institute (FSI) provides intensive language instruction to Foreign Service officers and formally assesses their language proficiency before they take on an assignment that requires the use of a language other than English.

To help FSI keep pace with current developments in language assessment, the agency asked the National Academies of Sciences, Engineering, and Medicine (the National Academies) to review the strengths and weaknesses of some key assessment[1] approaches that are available for assessing language proficiency[2] that FSI could apply in its context. The National Academies convened a committee of experts to conduct the review.

USING A PRINCIPLED APPROACH TO DEVELOP LANGUAGE ASSESSMENTS

The committee described a framework for developing and validating assessments that is depicted in Figure 1. The assessments and their use are central to test development and validation, but they are driven by foundational considerations related to the understanding of language, the contexts influencing the assessment, and the target language use that is the focus of the assessment. The committee calls this view a "principled approach" to assessment.

The essence of this approach is that specific choices for individual assessment methods and task types have to be understood and justified in the context of the specific ways that test scores are interpreted and used, rather than in the abstract: more consideration is required than a simple choice for an oral interview or a computer-adaptive reading test. The desirable technical characteristics of an assessment result from an iterative process that shapes key design and implementation decisions while considering evidence about how the decisions fit with the specific context in which they will be used.

Recent research in applied linguistics has led to a nuanced understanding of second and foreign language proficiency that goes well beyond a traditional focus on grammar and vocabulary. This newer perspective highlights the value of the expression of meanings implied in a given context, multiple varieties of any given language, the increasing use of multiple languages in a single conversation or context, and the recognition that communication in real-world settings typically uses multiple language skills in combination, frequently together with nonlinguistic modalities, such as graphics and new technologies.

Many of these more recent perspectives are relevant to the language needs of Foreign Service officers, who need to use the local language to participate in meetings and negotiations, understand broadcasts and print media, socialize informally, make formal presentations, and communicate using social media. The challenges presented by this complex range of Foreign Service tasks are reflected in the current FSI test and its long history of development.

1 Although in the testing field "assessment" generally suggests a broader range of approaches than "test," in the FSI context both terms are applicable, and they are used interchangeably in the report.

2 This report uses the term "language proficiency" to refer specifically to second and foreign language proficiency, which is sometimes referred to in the research literature as "SFL" or "L2" proficiency.

THE CURRENT FSI TEST

FSI's current test is given to several thousand State Department employees each year. It is given in 60 to 80 languages, with two-thirds of the tests in the five most widely used languages (Arabic, French, Mandarin Chinese, Russian, and Spanish). The assessment involves a set of verbal exchanges between the test taker and two evaluators: a "tester," who speaks the target language of the assessment and interacts with the test taker only in the target language, and an "examiner," who does not necessarily speak the target language and interacts with the test taker only in English.

The current test includes a speaking test and a reading test. The speaking test involves three parts: (1) conversations between the test taker and the tester about several different topics in the target language; (2) a brief presentation by the test taker to the tester, with follow-up questions; and (3) the test taker's interview of the tester about a specific topic, which is reported to the examiner in English. The reading test involves reading several types of material in the target language and reporting back to the examiner in English, including short passages for gist and longer passages in depth.

The tester and the examiner jointly determine the test taker's scores in speaking and reading through a deliberative, consensus-based procedure, considering and awarding points for five factors: comprehension, ability to organize thoughts, grammar, vocabulary, and fluency. The final reported scores are based on the proficiency levels defined by the Interagency Language Roundtable (ILR), a group that coordinates second and foreign language training and testing across the federal government. The ILR score levels are linked to personnel policies, including certification, job placement, retention in the Foreign Service, and pay.

POSSIBLE CHANGES TO THE FSI TEST

The committee considered possible changes to the FSI test that might be motivated by particular goals for improving it. Such goals might arise from an evaluation that suggests ways the current test should be strengthened. Table 1 summarizes the changes that the committee considered in terms of some potential goals for strengthening the current test. These are possible changes for FSI to consider further, not recommendations from the committee. Given its charge, the committee focused on possible changes that would address goals for improvement related to the construct assessed by the test, and the reliability

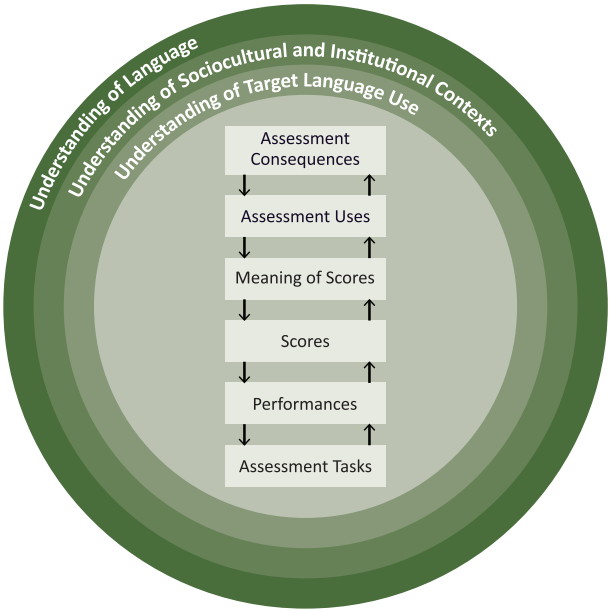


FIGURE 1 A principled approach to language assessment design and validation.

and fairness of its scores. The committee also noted instructional and practical considerations related to these possible changes.

BALANCING EVALUATION AND THE IMPLEMENTATION OF NEW APPROACHES

Central to FSI's considerations about how to strengthen its testing program lies a decision about the balance between (1) conducting an evaluation to understand how the current program is working and identifying changes that might be made in light of a principled approach to assessment design, and (2) starting to implement possible changes. Both are necessary for test improvement, but given limited time and resources, how much emphasis should FSI place on evaluation versus implementation?

Two questions can help inform this tradeoff:

First, does the FSI testing program have evidence related to the validity of the interpretation and use of test scores? For example:

- Comparisons of the specific language-related tasks carried out by Foreign Service officers with the specific language tasks on the FSI test
- Comparisons of the features of effective language use by Foreign Service officers in the field with the criteria that are used to score the FSI test
- Comparisons of the beliefs that test users have about the meaning of different FSI test scores

with the actual proficiency of Foreign Service officers who receive those scores

- Comparisons of the proficiency of Foreign Service officers in using the local languages to carry out typical tasks with the importance of those tasks to the job

Second, does the FSI testing program incorporate best practices related to the professional standards for workplace testing? For example:

- **Job Analyses** should be conducted regularly, typically every 3-5 years.
- **Content-Related Validity Evidence** should include documentation of the procedures used to determine the content and task formats on the test, and to develop and field-test the test questions.
- **Test Administration** instructions should be provided for administering the test under standard conditions, for providing testing accommodations to test takers who need them, and for maintaining the confidentiality and security of test content.
- **Scoring Processes** should be documented, including development of the scoring rubric and the procedures for scoring performances; for selecting, training, and re-training raters; for resolving score discrepancies; and for monitoring and evaluating the technical qualities of the scoring.
- **Cut-Score Setting** should be documented if performance or passing levels are defined with respect to the score scale, including details about the qualifications of panelists used to set cut scores, the instructions they were given, and their levels of agreement.
- **Psychometric and Other Technical Qualities** should be documented based on a program of research designed to evaluate reliability and decision consistency, rater agreement for the scoring of constructed-response questions, differential item functioning, and other relevant technical analyses, such as cognitive studies to understand how test takers process the test questions and corpus analyses to understand aspects of the target language use domain.
- **Efforts to Ensure Fairness** should be documented, including efforts to ensure that scores have the same meaning for all population groups in the intended testing population and that test takers with comparable proficiency receive comparable scores.
- **Score Reporting** processes should be documented, including procedures for determining the design and contents of the score reports.
- **The Purposes and Uses of the Test** should be documented.
- **An Explicit Statement of the Validity Argument** should be documented in an up-to-date technical report that is available to the public.
- **Information for Test Takers** should be provided to familiarize test takers with the test content and format.

If the answer to either of these questions is "no," it makes sense to place more weight on the evaluation side to better understand how the current program is working. If the answer to these questions is "yes," there is probably sufficient evidence to place more weight on the implementation side.

On the evaluation side, one important consideration is the institutional structure that supports research at FSI and provides an environment that allows continuous improvement. Many assessment programs incorporate regular input from researchers into the operation of their program, either from technical advisory groups or from visiting researchers and interns. Both of these routes allow assessment programs to receive new ideas from experts who understand the testing program and can provide tailored advice.

On the implementation side, options for making changes may be constrained by two long-standing FSI policies:

- Assessing all languages with the same approach: the desire for comparability that underlies this policy is understandable, but what is essential is the comparability of results from the test, not the comparability of the testing processes.
- The use of the ILR framework: the ILR framework is useful for coordinating personnel policies across government agencies, but that does not mean it has to be used for all aspects of the FSI test.

These two policies may be more flexible than it might seem, so FSI may have substantially more opportunity for innovation and continuous improvement in its testing program than has been generally assumed.

TABLE 1 Possible Changes to the FSI Test to Meet Potential Goals

Possible Change	Potential Test Construct, Reliability and Fairness Considerations	Potential Instructional and Practical Considerations
Using multiple measures	Better coverage of Foreign Service language uses Greater reliability and fairness	Additional cost for test development and administration
Scoring listening on the speaking test	More systematic use of listening information already generated by the test Possibility of increased measurement error	Potential for positive effect on instruction Additional complexity to the scoring process
Adding target-language writing as a response mode for some reading or listening tasks	Coverage of Foreign Service language uses that involve writing	Potential for positive effect on instruction Extra cost for test development and administration
Adding paired or group oral tests	Better coverage of Foreign Service language uses related to interactional competence Possibility of increased measurement error due to partner variability	Potential for positive effect on instruction Cost and practical challenges of coordinating tests
Using recorded listening tasks that use a range of language varieties and unscripted texts	Potential for better generalization of listening assessment to typical range of Foreign Service contexts	Potential for positive effect on instruction Increased cost for test development and administration
Incorporating language supports (such as dictionary and translation apps)	Better coverage of Foreign Service language uses	Minor modifications to current test
Adding a scenario-based assessment	Better coverage of complex Foreign Service language uses	Potential for positive effect on instruction Increased cost for test development and administration
Incorporating portfolios of work samples	Better coverage of Foreign Service language uses Potential for increased overall reliability and fairness by using multiple measures	Difficult to standardize Extra cost for development of scoring criteria and procedures
Adding computer-administered tests using short tasks in reading and listening	Better coverage and reliability for Foreign Service professional topics	Additional cost and administrative steps, which may be prohibitive for low-volume languages
Using automated assessment of speaking	Potential to increase standardization	Capabilities are limited but improving Future potential to decrease cost of test administration Expensive to develop, so cost-effective only for high-volume tests
Providing transparent scoring criteria	Potential for greater reliability and fairness	Minor modifications of current test information procedures
Using additional scorers	Potential for greater reliability and fairness	Minor modification of current test procedures Additional cost
Providing more detailed score reports	Better understanding of scores for all users of FSI test	Potential for positive effect on instruction Increased cost and time for score reporting

For More Information . . . This Consensus Study Report Highlights was prepared by the Committee on National Statistics based on the Consensus Study Report, *A Principled Approach to Language Assessment: Considerations for the U.S. Foreign Service Institute* (2020). The study was sponsored by the U.S. Foreign Service Institute. Any opinions, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect the views of any organization or agency that provided support for the project. Copies of the Consensus Study Report are available from the National Academies Press, (800) 624-6242; http://www.nas.edu/25748

COMMITTEE ON FOREIGN LANGUAGE ASSESSMENT FOR THE U.S. FOREIGN SERVICE INSTITUTE

DORRY M. KENYON (*Chair*), Center for Applied Linguistics; **DAVID DORSEY**, HumPRO; **LORENA LLOSA,** New York University; **ROBERT J. MISLEVY,** Educational Testig Service; **LIA PLAKANS**, University of Iowa; **JAMES PURPURA,** Columbia University; **M. ELVIS WAGNER,** Temple University; **PAULA M. WINKE**, Michigan State University; **STUART ELLIOTT,** *Study Director;* **JUDITH KOENIG**, *Senior Program Officer;* **NATALIE NIELSEN,** *Senior Program Officer;* **ANTHONY MANN,** *Program Associate*.