A PRINCIPLED APPROACH TO LANGUAGE ASSESSMENT

Considerations for the U.S. Foreign Service Institute

Committee on Foreign Language Assessment for
the U.S. Foreign Service Institute

Division of Behavioral and Social Sciences and Education

A Consensus Study Report of

The National Academies of
SCIENCES · ENGINEERING · MEDICINE

THE NATIONAL ACADEMIES PRESS
Washington, DC
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001

This activity was supported by contracts between the National Academy of Sciences and the U.S. Department of State Foreign Service Institute (award #19FS1A18C0002). 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.

International Standard Book Number-13: 978-0-309-67548-2
International Standard Book Number-10: 0-309-67548-0
Digital Object Identifier: https://doi.org/10.17226/25748

Additional copies of this publication are available from the National Academies Press, 500 Fifth Street, NW, Keck 360, Washington, DC 20001; (800) 624-6242 or (202) 334-3313; http://www.nap.edu.

Printed in the United States of America

Suggested citation: National Academies of Sciences, Engineering, and Medicine. (2020). *A Principled Approach to Language Assessment: Considerations for the U.S. Foreign Service Institute*. Washington, DC: The National Academies Press. doi: https://doi.org/10.17226/25748.

The National Academies of

SCIENCES • ENGINEERING • MEDICINE

The **National Academy of Sciences** was established in 1863 by an Act of Congress, signed by President Lincoln, as a private, nongovernmental institution to advise the nation on issues related to science and technology. Members are elected by their peers for outstanding contributions to research. Dr. Marcia McNutt is president.

The **National Academy of Engineering** was established in 1964 under the charter of the National Academy of Sciences to bring the practices of engineering to advising the nation. Members are elected by their peers for extraordinary contributions to engineering. Dr. John L. Anderson is president.

The **National Academy of Medicine** (formerly the Institute of Medicine) was established in 1970 under the charter of the National Academy of Sciences to advise the nation on medical and health issues. Members are elected by their peers for distinguished contributions to medicine and health. Dr. Victor J. Dzau is president.

The three Academies work together as the **National Academies of Sciences, Engineering, and Medicine** to provide independent, objective analysis and advice to the nation and conduct other activities to solve complex problems and inform public policy decisions. The National Academies also encourage education and research, recognize outstanding contributions to knowledge, and increase public understanding in matters of science, engineering, and medicine.

Learn more about the National Academies of Sciences, Engineering, and Medicine at **www.nationalacademies.org.**

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

Consensus Study Reports published by the National Academies of Sciences, Engineering, and Medicine document the evidence-based consensus on the study's statement of task by an authoring committee of experts. Reports typically include findings, conclusions, and recommendations based on information gathered by the committee and the committee's deliberations. Each report has been subjected to a rigorous and independent peer-review process and it represents the position of the National Academies on the statement of task.

Proceedings published by the National Academies of Sciences, Engineering, and Medicine chronicle the presentations and discussions at a workshop, symposium, or other event convened by the National Academies. The statements and opinions contained in proceedings are those of the participants and are not endorsed by other participants, the planning committee, or the National Academies.

For information about other products and activities of the National Academies, please visit www.nationalacademies.org/about/whatwedo.

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

DORRY M. KENYON (*Chair*), Center for Applied Linguistics, Washington, DC
DAVID DORSEY, Human Resources Research Organization, Alexandria, VA
LORENA LLOSA, Department of Teaching and Learning, New York University
ROBERT J. MISLEVY, Educational Testing Service, Princeton, NJ
LIA PLAKANS, College of Education, University of Iowa
JAMES E. PURPURA, Teachers College, Columbia University
M. "ELVIS" WAGNER, College of Education, Temple University
PAULA M. WINKE, Department of Linguistics and Languages, Michigan State University

STUART W. ELLIOTT, *Study Director*
JUDITH KOENIG, *Senior Program Officer*
NATALIE NIELSEN, *Consultant*
ANTHONY S. MANN, *Program Associate*

Preface

This project resulted from a request of the U.S. Foreign Service Institute (FSI) to the National Academies of Sciences, Engineering, and Medicine to provide input related to the assessment of the language proficiency of Foreign Service personnel. Throughout the study, the committee was guided by its interactions with representatives of FSI, who explained the details and the context of FSI's current assessment as well as their goals for the study. In particular, we held three extended discussions with representatives of FSI's School of Language Studies, led by Ambassador Wanda Nesbitt, dean; James North, associate dean for instruction; David Sawyer, director, Language Testing Unit; and Benjamin Kloda, evaluation coordinator. We also appreciate Dr. Sawyer's facilitation for members of the committee to take the current FSI assessment: some took the speaking test remotely, and some took the full test onsite.

In the course of planning the project and identifying prospective members of the committee, the National Academies received input from a wide range of researchers in language assessment and related fields. For their advice and insights during the early stages of the project, we thank the many individuals who helped us: Randy Bennett, Educational Testing Service; Rachel Brooks, Federal Bureau of Investigation; Carol A. Chapelle, Department of English, Iowa State University; Alister Cumming, Ontario Institute for Studies in Education, University of Toronto; Sara Cushing, Department of Applied Linguistics and English as a Second Language, Georgia State University; Steve Ferrara, Measured Progress, Inc.; Neus Figueras, University of Barcelona; Glenn Fulcher, Department of English, University of Leicester; Luke Harding, Department of Linguistics

and English Language, Lancaster University; Okim Kang, Department of Applied Linguistics, Northern Arizona University; YouJin Kim, Department of Applied Linguistics and English as a Second Language, Georgia State University; Deirdre Knapp, Human Resources Research Organization; Antony John Kunnan, Department of English, University of Macau; Patrick Kyllonen, Educational Testing Service; Beth A. Mackey, National Cryptologic School, Central Intelligence Agency; Margaret E. Malone, American Council on the Teaching of Foreign Languages; Rodney A. McCloy, Human Resources Research Organization; John Norris, Educational Testing Service; Gary Ockey, Linguistics Program, Iowa State University; Lourdes Ortega, Department of Linguistics, Georgetown University; Frederick L. Oswald, Department of Psychological Sciences, Rice University; Carsten Roever, School of Languages and Linguistics, University of Melbourne; Steven J. Ross, School of Languages, Literatures, and Cultures, University of Maryland, College Park; Sun-Young Shin, Department of Second Language Studies, Indiana University, Bloomington; Xiaoming Xi, Educational Testing Service; and Rebecca Zwick, Educational Testing Service.

This Consensus Study Report was reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published report as sound as possible and to make certain that the report meets institutional standards for objectivity, evidence, and responsiveness to the charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process.

We thank the following individuals for their review of this report: Carol A. Chapelle, Applied Linguistics Program, Department of English, Iowa State University; Brian E. Clauser, Measurement Consulting Services, National Board of Medical Examiners; Alister Cumming, Centre for Educational Research on Languages and Literacies, Ontario Institute for Studies in Education, University of Toronto; Luke Harding, Department of Linguistics and English Language, Lancaster University; Okim Kang, Applied Linguistics Speech Lab, Northern Arizona University; Patricia K. Kuhl, Institute for Learning and Brain Sciences, University of Washington; Margaret E. Malone, Assessment, Research and Development, American Council on the Teaching of Foreign Languages; Frederick L. Oswald, Department of Psychological Sciences, Rice University; and Steven J. Ross, School of Languages, Literatures, and Cultures, University of Maryland, College Park.

Although the reviewers listed above provided many constructive comments and suggestions, they were not asked to endorse the content of the report nor did they see the final draft of the report before its release. The

review of this report was overseen by Lorrie A. Shepard, Research and Evaluation Methodology, School of Education, University of Colorado Boulder, and Eugenie C. Scott, executive director (retired), National Center for Science Education. They were responsible for making certain that an independent examination of this report was carried out in accordance with the standards of the National Academies and that all review comments were carefully considered. Responsibility for the final content rests entirely with the authoring committee and the National Academies.

Throughout this project, I have had the privilege to work with the sterling group of colleagues who served as fellow members on the committee. All of them enthusiastically took time from their many professional commitments to work together to understand FSI's testing program and consider how to present and highlight the important and relevant research and practice from the field of language testing. During our deliberations, the members were often reminded that our goal for the report was to distill the messages from the research literature in the field of language assessment into a form that we could discuss over lunch with our colleagues from FSI after the project concluded.

Our four meetings together were unfailingly intense and productive, with everyone contributing to advancing our common understanding and testing each other's arguments. Between meetings, every member tirelessly and cheerfully drafted and critiqued text, tracked down details, and clarified points. I am also grateful for the support of the staff throughout the project, to provide the committee with a supportive environment for our deliberations and to challenge us to clarify our messages for FSI. It has been a great pleasure for me to work with such a wonderful group of committee members and staff over the course of this study.

In carrying out this project, the committee was impressed by FSI's sensitive appreciation of the issues related to language testing and the agency's desire to foster long-term improvement in its language testing program. We hope this report inspires FSI with a sense of opportunity and direction for its work ahead.

Dorry Kenyon, *Chair*
Committee on Foreign Language
Assessment for the U.S. Foreign
Service Institute

Contents

Summary

The U.S. Department of State 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. The State Department uses the results of the FSI assessment to make decisions related to certification, job placement, promotion, retention, and pay.

To help FSI keep pace with current developments in language assessment, the agency asked the National Academies of Sciences, Engineering, and Medicine to conduct a review of 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. FSI requested a report that provides considerations about relevant assessment approaches without making specific recommendations about the approaches the agency should adopt

[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 throughout this 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 report does not address the assessment of language proficiency of native speakers (e.g., as in an assessment of the reading or writing proficiency of U.S. high school students in English) except in the case of native speakers of languages other than English who need to certify their language proficiency in FSI's testing program.

and without evaluating the agency's current testing program. This request included an examination of important technical characteristics of different assessment approaches. The National Academies formed the Committee on Foreign Language Assessment for the U.S. Foreign Service Institute to conduct the review.

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 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. The committee calls this view a "principled approach" to assessment.

USING A PRINCIPLED APPROACH TO DEVELOP LANGUAGE ASSESSMENTS

The considerations involved in developing and validating language assessments and the ways they relate to each other are shown in Figure S-1. The assessment and its use are in the center of the figure, with the boxes and arrows describing the processes of test development and validation. Surrounding the assessment and its use are the foundational considerations that guide language test development and validation: the understanding of language, the contexts influencing the assessment, and the target language use that is the focus of the assessment.

A principled approach to language test development explicitly takes all these factors into account, using evidence about them to develop and validate a test. In particular, a principled approach considers evidence in two complementary ways: (1) evidence that is collected as part of the test about the test takers to support inferences about their language proficiency, and (2) evidence that is collected about the test and its context to evaluate the validity of its use and improve the test over time.

FOUNDATIONAL CONSIDERATIONS

One key aspect of a principled approach to developing language assessments involves the understanding of how the target language is used in real life and how that use motivates the assessment of language proficiency. This understanding is crucial not only for initial test development, but also for evaluating the validity of the interpretations and uses of test results and for improving a test over time. There are a number of techniques for analyzing

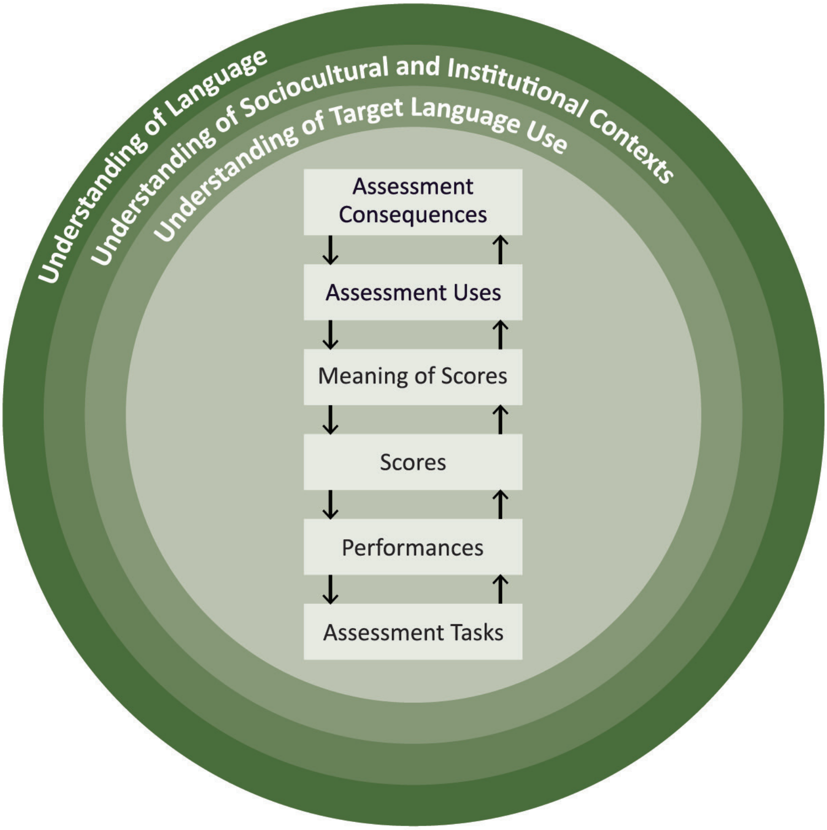
Understanding of Language
Understanding of Sociocultural and Institutional Contexts
Understanding of Target Language Use
Assessment Consequences
Assessment Uses
Meaning of Scores
Scores
Performances
Assessment Tasks

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

language use in a domain that could be used to refine FSI's current understanding of language use in the Foreign Service context.

Research in applied linguistics over the past few decades 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 on language proficiency 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.

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 test includes two parts: a speaking test and a reading test. The speaking test involves (1) conversations between the test taker and the tester about several different topics in the target language; (2) a brief introductory statement 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—short passages for gist and longer passages in depth—and reporting back to the examiner in English, responding to follow-up questions from the examiner or the tester as requested.

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 level scores 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 in response to particular goals for improving the test. Such goals might arise from an evaluation of the validity of the interpretations and uses of the test, guided by a principled approach, which suggests particular ways the current test should be strengthened. Table S-1 summarizes changes that

TABLE S-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

continued

TABLE S-1 Continued

Possible Change	Potential Test Construct, Reliability and Fairness Considerations	Potential Instructional and Practical Considerations
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 • 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

the committee considered for the FSI test in terms of some potential goals for strengthening the current test. Given its charge, the committee specifically focused on possible changes that would address goals for improvement related to the construct assessed by the test, and the reliability and fairness of its scores. In addition, the committee noted potential instructional and practical considerations related to these possible changes.

CONSIDERATIONS IN EVALUATING VALIDITY

Evaluating the validity of the interpretation and use of test scores is central to a principled approach to test development and use. Such evaluations consider many different aspects of the test, its use, and its context.

Several kinds of evidence could be key parts of an evaluation of the validity of using FSI's current test:

- 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; and
- 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.

As a "high-stakes" test—one that is used to make consequential decisions about individual test takers—it is especially important that the FSI test adhere to well-recognized professional test standards. One key aspect of professional standards is the importance of careful and systematic documentation of the design, administration, and scoring of a test as a good practice to help ensure the validity, reliability, and fairness of the interpretations and decisions supported by a testing program.

BALANCING EVALUATION AND THE IMPLEMENTATION OF NEW APPROACHES

At the heart of the FSI's choice 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 each?

Two questions can help address this tradeoff:

1. Does the FSI testing program have evidence related to the four example comparisons listed above?
2. Does the program incorporate the best practices recommended by various professional standards?

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:

1. 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.
2. 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.

Complicated choices will need to be made about how to use a principled approach to assessment, select which language assessment options to explore, and set the balance between evaluation and implementation. In requesting this report, FSI has clearly chosen a forward-looking strategy. Using this report as a starting point and thinking deliberatively about these complicated choices, FSI could enhance its assessment practices by improving its understanding of the test construct and how it is assessed; the reliability of the test scores and the fairness of their use; the potential beneficial influence of the test on instruction; and the understanding, usefulness, and acceptance of the assessment across the State Department community.

1

Introduction

The United States is formally represented around the world by approximately 14,000 Foreign Service officers and other personnel in the U.S. Department of State. Roughly one-third of them are required to be proficient in the local languages of the countries to which they are posted. To achieve this language proficiency for its staff, the State Department's Foreign Service Institute (FSI) provides intensive language instruction and assesses the proficiency of personnel before they are posted to a foreign country. The requirement for language proficiency is established in law and is incorporated in personnel decisions related to job placement, promotion, retention, and pay. FSI also tests the language proficiency of the spouses of Foreign Service officers, as a point of information, as well as Foreign Service personnel from other U.S. government agencies.

BACKGROUND

Given recent developments in language assessment, FSI asked the National Academies of Sciences, Engineering, and Medicine to review the strengths and weaknesses of key assessment[1] approaches for assessing lan-

[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 throughout this report.

guage proficiency[2] that would be relevant for its language test. In response, the National Academies formed the Committee on Foreign Language Assessment for the U.S. Foreign Service Institute to conduct the review; Box 1-1 contains the committee's statement of task.

FSI's request was motivated by several considerations. First, although FSI's assessment has been incrementally revised since it was developed in the 1950s, significant innovations in language assessment since that time go well beyond these revisions. Examples include the use of more complex or authentic assessment tasks, different applications of technology, and the collection of portfolios of language performances over different school or work settings. Second, in the FSI environment, questions have arisen about limitations or potential biases associated with the current testing program. Third, the nature of diplomacy and thus the work of Foreign Service officers have changed significantly in recent decades. These changes mean that the language skills required in embassy and consulate postings are different from those needed when the FSI test was developed. For example, transactions that once took place in person are now often conducted over email or by text, and television and the Internet are increasingly prominent sources of information. For these reasons, FSI wanted to take a fresh look at language testing options that are now available and that could be relevant to testing the language proficiency of Foreign Service officers.

THE COMMITTEE'S APPROACH

The committee's charge includes questions about specific approaches to language assessment and their psychometric characteristics. In addressing this charge, the committee began from the fundamental position that choices for assessment methods and task types have to be understood and justified in the context of the ways that test scores are interpreted and used, not abstractly. Also, concerns about fairness, reliability, and other psychometric characteristics should be addressed through the evaluation of the validity of an assessment for its intended use, not abstractly.

Thus, the committee began its deliberations by considering relatively new approaches for designing and developing language assessments that have been growing in use in the measurement field. These approaches are referred to as "principled" because they are grounded in the principles of evidentiary reasoning (see Mislevy and Haertel, 2006; National Research

[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 report does not address the assessment of language proficiency of native speakers (e.g., as in an assessment of the reading or writing proficiency of U.S. high school students in English) except in the case of native speakers of languages other than English who need to certify their language proficiency in FSI's testing program.

BOX 1-1
Statement of Task

An ad hoc committee will evaluate the different approaches that (1) exist to assess foreign language proficiency and that (2) the State Department's Foreign Service Institute (FSI) could potentially use to assess language proficiency. The committee will consider the key assessment approaches in the research literature that are appropriate for language testing, including, but not limited to, assessments that use task-based or performance-based approaches, adaptive online test administration, and portfolios.

The committee will collect information that helps answer the following questions:

- What assessment formats and approaches are feasible for language proficiency testing? What are the advantages and disadvantages of the various approaches?
- How well do different assessment approaches measure reading and listening comprehension (interactive and non-interactive)?
- How well do different assessment approaches measure speaking proficiency?
- To what extent would different assessment approaches provide information to support the intended inferences about a candidate's language proficiency?
- What are the psychometric characteristics (reliability, validity, classification accuracy) associated with different approaches?
- Are the different assessment approaches equally effective (fair and unbiased) for all groups that typically take the FSI assessments?
- To what extent is unconscious bias a concern with different assessment strategies? Which assessment approaches minimize the effect of unconscious bias in foreign language proficiency testing?
- Are the different assessment approaches equally practical and cost effective in a resource-limited government environment?
- The committee will not recommend any specific assessment approach but will describe the strengths and weaknesses of different assessment approaches, in light of the latest research and current scientific consensus. The committee will also take into account the practicality of various options in a resource-limited, government environment (in contrast to academic or private sector assessment applications). To the extent possible, the study should address the steps involved in conducting proficiency assessments to ultimately enable the State Department to determine the most appropriate method to utilize for the Foreign Service.
- The assessment process currently used by FSI and the definition of language proficiency developed by the Interagency Language Roundtable provide the context for the study. However, the purpose of the consensus study is not to evaluate FSI's current assessment process. That process can serve as one possible benchmark for comparison when identifying the strengths and weaknesses of other assessment approaches. The focus of the study is also not to evaluate the current definition of language proficiency used by FSI, or its approach to language learning, but instead to identify the most effective means of assessing language proficiency as currently defined in the context of the U.S. Foreign Service.

Council, 2001b, 2014). They begin with the context and intended use of scores for decision making and then assemble evidence and build an argument to show that the assessment provides reliable and valid information to make those decisions. The committee judged that these approaches would be useful for informing potential changes to the FSI test.

Likewise, the committee focused on FSI's intended uses of the test to shape its review of the literature on assessment methods. Rather than starting with specific methods or task types—such as assessments that use task-based or performance-based methods, online test administration, or portfolios—the committee focused on the types of information that different methods might yield and how well that information aligned with intended uses. The committee's analyses are in no way exhaustive because the charge did not include redesigning FSI's assessment system. Instead, the committee identified some potential goals for strengthening the test and then considered the changes that might help achieve those goals.

FSI's intended use specifically relates to a set of job-related tasks that are done by Foreign Service officers. To better understand the kinds of assessment methods that might be most relevant to this use, the committee sought information about the language tasks that officers perform on the job and the nature of the decisions that need to be made about test takers based on their language proficiency. Thus, one of the most important aspects of the committee's information gathering was a series of discussions with FSI representatives about the current context and practice of language assessment in the agency. These discussions provided an analytical lens for the committee's literature review. In addition, FSI provided an opportunity for many of the committee members to take the current test themselves.

In its review of the literature, the committee was also strongly influenced by the evolution of the understanding of language use and the implications of that understanding for the design of assessments. These trends have heightened the appreciation of several common features of second language use, including the need to evaluate how examinees use language to communicate meanings, the integrated use of language across modalities, and the prevalence of multilingualism in many situations in which multiple languages are used.

A PRINCIPLED APPROACH TO ASSESSMENT DESIGN AND VALIDATION

In this section, the committee provides a framework for thinking about how to develop, implement, and administer an assessment, and then monitor and evaluate the validity of its results: in FSI's case an assessment of language proficiency for Foreign Service officers. This framework is built on a "principled approach" to assessment design and validation. The section begins with a review of important measurement properties, followed by a more detailed explanation of what a principled approach involves.

Fundamental Measurement Properties

Assessment, at its core, is a process of reasoning from evidence. The evidence comes from a test-taker's performance on a set of assessment tasks. This performance serves as the basis for making inferences about the test-takers' proficiency in relation to the broader set of knowledge, skills, abilities, and other characteristics that are needed to perform the job and are the focus of the assessment. The process of collecting evidence to make inferences about what test takers know and can do is fundamental to all assessments. A test design and evaluation process seeks to ensure the quality of those inferences.

Validity is the paramount measurement property. Validity refers to "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (American Educational Research Association et al., 2014, p. 11). Validation is the process of accumulating evidence to provide a sound basis for proposed interpretations and uses of the scores.

Reliability refers to the precision of the test scores. Reliability reflects the extent to which the scores remain consistent across different replications of the same testing procedure. Reliability is often evaluated empirically using the test data to calculate quantitative indices. There are several types of indices, and each provides a different kind of information about the precision of the scores, such as the extent to which scores remain consistent across independent testing sessions, across different assessment tasks, or across raters or examiners. These indices estimate reliability in relation to different factors, one factor at a time. Other approaches for looking at the consistency of test scores—referred to as "generalizability analyses"—can estimate the combined effect of these different factors (see Mislevy, 2018; Shavelson and Webb, 1991).

Fairness in the context of assessment covers a wide range of issues. The committee addresses it under the broad umbrella of validity, focusing on the validity of intended score interpretations and uses for individuals from different population groups. Ideally, scores obtained under the specified test administrations procedures should have the same meaning for all test takers in the intended testing population.

The Concept of a Principled Approach to Test Development

Recent decades have seen an increasing use of approaches in general assessment that have become known as "principled" (Ferrara et al., 2017). A principled approach relies on evidentiary reasoning to connect the various pieces of assessment design and use. That is, it explicitly connects the design for assessment *tasks* to the *performances* that those tasks are intended to elicit from a test taker, to the *scores* derived from those performances, and to the *meaning* of those scores that informs their *use*. A principled approach specifically focuses on "validity arguments that support intended score interpretations and uses and development of empirical evidence to support those score interpretations and uses throughout the design, development, and implementation process" (Ferrara et al., 2016, p. 42).

The use of a principled approach to test development is intended to improve the design of assessments and their use so that the inferences based on test scores are valid, reliable, and fair. The foundations behind the use of a principled approach are detailed in *Knowing What Students Know* (National Research Council, 2001b, Ch. 5). Most notably, the use of a principled approach intertwines models of cognition and learning with models of measurement, which govern the design of assessments.

Figure 1-1, developed by the committee using ideas discussed by Bachman (2005), Bachman and Palmer (2010) and Kelly et al. (2018), depicts a set of key considerations involved in language assessment design that are emphasized by a principled approach. The assessment and its use are in the center of the figure, with the boxes and arrows illustrating the logic of test development and validation. Although the arrows suggest a rough ordering of the processes, they are inevitably iterative as ideas are tried, tested, and revised. Surrounding the test and its use, the figure shows the foundational considerations that guide language test development in three rings. Specifically, these rings reflect the test developer's understanding of language, the sociocultural and institutional contexts for the assessment, and the target language use domain that is the focus of the assessment.

Developed in light of these foundational considerations, the assessment contains tasks that elicit performances, which are evaluated to produce scores, which are interpreted as an indicator of proficiency, which are then used in various ways. The decision to use these scores for any given purpose carries consequences for the test takers and many others. This chain of relationships is fundamental to understanding the design of an assessment and the validity of interpreting its results as a reflection of test-takers' language proficiency. The validation process collects evidence about these various relationships.

Figure 1-1 is used to structure the report. Chapter 2 describes the FSI context and current test, which reflects all aspects of the figure. Chapter 3 addresses a set of relevant concepts and techniques for understanding the three rings surrounding the test. Chapter 4 addresses possible changes to the tasks, performances, and scoring of the current FSI test. Chapter 5 addresses considerations related to the meaning of the test scores and the way

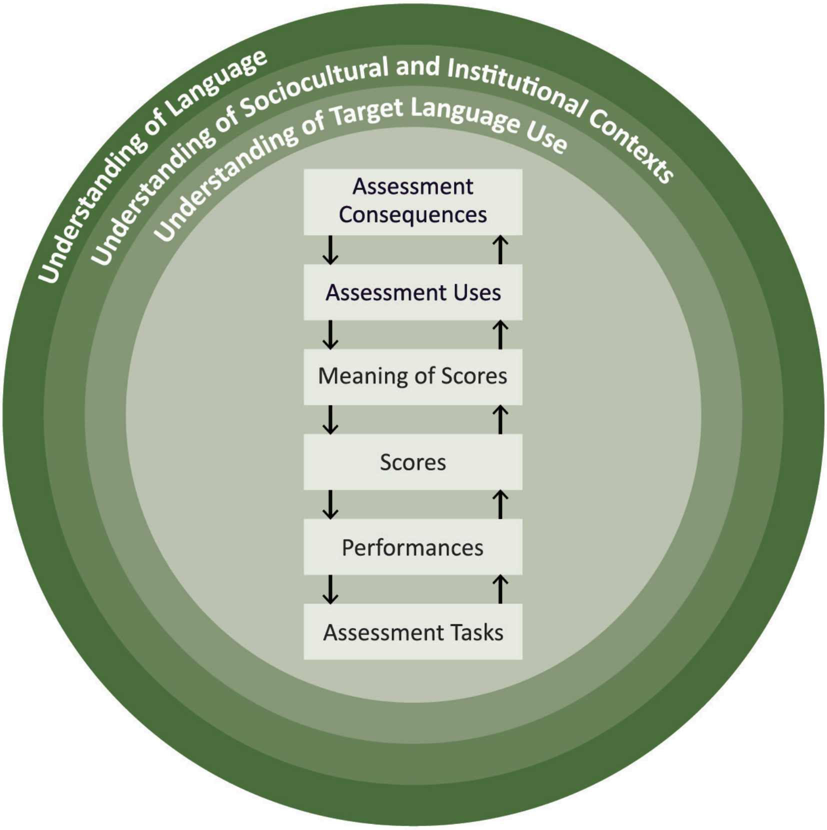
Understanding of Language
Understanding of Sociocultural and Institutional Contexts
Understanding of Target Language Use
Assessment Consequences
Assessment Uses
Meaning of Scores
Scores
Performances
Assessment Tasks

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

they map onto uses and consequences. Chapter 6 then discusses the validity arguments that concern the relationships of the elements in the figure. Finally, the report closes by considering how to balance limited time and resources between evaluation to understand how the current test is doing and the implementation of new approaches.

FSI's request to the National Academies was framed in the context of FSI's testing needs, and the field's principled approaches specifically direct attention toward the context and intended use for an assessment. As a result, many of the details of the report are necessarily geared toward the context of FSI's language assessment needs. Despite this focus, however, the committee hopes the report will be useful to other organizations with language testing programs, for both government language testing and the larger community. The lessons related to the need for building from a clear argument within the context of test use to assessment design are applicable to all testing programs, even if the specifics of the discussion in this report relate primarily to FSI. In addition, the range of possible design choices for an assessment program are similar across programs, even if the specific contexts of different programs will affect which of those choices may be most appropriate.

2

The FSI Testing Context

This chapter provides an overview of the FSI testing context. It discusses how languages are used in the Foreign Service (the target language use domain of the test), how FSI assesses the language proficiency of Foreign Service officers and other Foreign Service personnel, and how the results of those assessments are used in the State Department.

MANDATE FOR ASSESSING FOREIGN LANGUAGE PROFICIENCY

Foreign Service officers are posted to nearly every country in the world. As of December 2018, the U.S. State Department was operating 170 embassies and 107 consulates or missions to international organizations. In some countries, such as Brazil and China, the United States operates an embassy and several consulates.

Most U.S. embassies include eight job categories that require the greatest use of foreign language. The job categories and their associated language uses are summarized in Table 2-1.

Foreign Service officers are expected to be able to function effectively and professionally in these capacities. Accordingly, their language proficiency is expected to be adequate to perform in the local language across the job categories. For this reason, foreign language proficiency is a central feature in the professional development of U.S. diplomats and is required for many Foreign Service officers. This requirement was established in the Foreign Service Act of 1980, as amended, and has been incorporated into high-stakes personnel decisions relating to tenure, promotion, overseas

TABLE 2-1 Summary of U.S. Embassy Job Categories and Language Uses

Job Category	General Roles and Responsibilities	Broad Language Uses
Chief of Mission (the ambassador)[a]	• Lead the embassy	• Explain and garner support for U.S. foreign policy • Explain American customs, values, and traditions
Deputy Chief of Mission	• Manage the mission • Facilitate interagency coordination • Act as chargé d'affaires in the ambassador's absence	• Speak about any topic the ambassador would address
Public Affairs Officers	• Liaise with the media • Organize cultural and educational exchanges • Oversee embassy's social media accounts • Organize programs for American speakers in the host country	• Provide information about American policy and society • Interview and select participants for exchange programs
Political Officers	• Analyze the political climate	• Understand all local and regional political and economic developments • Report to Washington headquarters about potential effects on U.S. foreign policy and U.S. interests
Economic Officers	• Track and analyze the host country and region's economy, and the economic effect of U.S. policies in the host country	• Gather and analyze economic and political information • Report to Washington headquarters about the host country's economy and politics
Consular Officers	• Adjudicate visas • Provide wide range of services to U.S. citizens	• Interview visa applicants • Interact with wide range of local entities, such as the legal system or hospitals
Management Officers	• Manage the embassy's or consulate's administrative support services	• Interact with a variety of local government officials, organizations, and individuals on a wide range of issues
Regional Security Officers	• Ensure the safety and security of people and facilities	• Interact with a variety of local government officials, organizations, and individuals on a wide range of issues

[a] The broad language uses for chiefs of mission apply generally to career Foreign Service officers serving as ambassadors, although language proficiency is not a prerequisite for an appointment. Career Foreign Service officers who serve as chiefs of mission usually acquire relevant language skills during the course of their careers. However, most political appointees serving as ambassadors receive little or even no language training immediately prior to assignment as a chief of mission and conduct business in English or through a translator or interpreter.

postings, and incentive pay. The law directs the Secretary of State to establish foreign language proficiency requirements though it does not prescribe how the secretary should define or measure proficiency (Foreign Service Act of 1980, as amended, Section 702, 22 U.S.C 4022 et seq.):

> The Secretary of State shall establish foreign language proficiency requirements for members of the service who are to be assigned abroad in order that Foreign Service posts abroad will be staffed by individuals having a useful knowledge of the language or dialect common to the country in which the post is located.

LANGUAGE NEEDS AND TRAINING OF FOREIGN SERVICE OFFICERS

FSI's School of Language Studies provides intensive language training on a full-time basis for Foreign Service officers to develop their language proficiency. The FSI school is also used by other U.S. government agencies that assign personnel abroad, including the Agency for International Development and the U.S. Department of Defense. The school provides instruction in more than 65 languages. The length of training depends on the difficulty of the language for English speakers, ranging from 24 to 30 weeks (Spanish, French) to 88 weeks (Arabic, Japanese).

Every year, FSI surveys State Department employees who completed FSI language training during the previous fiscal year and who are currently serving in a language-designated position. The survey asks them how they are using their language skills in their work and how well FSI language training prepared them to do so.

Aggregated results from the 2012 to 2016 surveys show a need to frequently perform a wide range of activities in the local language related to their jobs. As reported on the surveys, the most commonly used language activities include

- socializing both informally and in business settings,
- understanding meeting discussions and social conversations,
- understanding job-related documents,
- understanding broadcast and print media,
- communicating over the telephone and through e-mail,
- interviewing to elicit information,
- making presentations,
- writing social correspondence,
- giving instructions or explaining procedures, and
- monitoring and interacting using social media.

In some situations, locally employed staff who are native speakers of the local language can assist, but in regular everyday settings some level of language proficiency by the Foreign Service officers is essential. Although not addressed in the survey, it is likely that Foreign Service officers carry out these various tasks with some assistance from language supports, including dictionary and translation apps.

The goal of the language training is to prepare Foreign Service officers to participate effectively in this wide range of language activities. The focus is on a level of professional language proficiency that would allow Foreign Service officers to carry out any of the formal or informal job activities associated with language-designated positions in embassies and consulates. Box 2-1 illustrates the range of these activities with examples of

BOX 2-1
Examples of Critical Language Uses by Foreign Service Officers

Consular Officer: [During] language training, we practiced taking calls regarding welfare and whereabouts cases. We practiced using our mobile phones and via sometimes poor connections. At the time, I admit it felt somewhat forced, but at post I realized how especially valuable that practice would be to my job. I work in American Citizen Services and have several times been required to take calls regarding missing persons or persons in the hospital. . . . In multiple instances, I have acted as liaison and translator for distraught families.

Information Officer: I recently took a call from a source describing a potential attack against a border post we support. It was very helpful to speak enough [language] to understand what the source was saying.

Labor/Political Officer: My reading skills have been invaluable since I arrived at post 6 months ago. I read [language] news almost every other day, and it is particularly helpful when reading hard-copy newspapers, which are often the best way to obtain information on [country].

Political/Economic Officer: In my first month at post, I read a 500+ page parliamentary report on the [city] terrorist attacks in very legal/technical language that was only available in [local languages 1 and 2].

Regional Security Officer: I use my language every day to speak to local guard staff. Recently, I was able to learn about an employment issue that affected the morale of the guards. It was only my ability to speak with the guards casually that allowed me to learn of the issue. It would never have been brought to my attention by their immediate supervisors otherwise. It is satisfying to be able to have simple conversations that can lead to more substantive issues.

anonymized responses to a question on the Annual Language Impact Survey that asked respondents to describe a memorable time when they used their language skills effectively on the job.

CURRENT FSI TESTING

Overview

In fiscal 2018, FSI directly administered 3,364 tests in 63 languages to Foreign Service officers and other government agency personnel, and it outsourced 802 tests for external candidates for limited career appointments in consular affairs at overseas posts. This test volume is generally as it has been in recent years, although the volume has decreased since peaking at 5,729 in 2011. The number of languages tested in fiscal 2018 also was slightly lower than had been usual over the past decade, when approximately 80 languages were tested each year.

About two-thirds of the tests are in five widely used languages: Arabic (260), French (583), Mandarin Chinese (271), Russian (208), and Spanish (1,071). The tests in the remaining languages are given to far fewer people, including 35 languages with 10 or fewer test takers. (All data are for fiscal 2018 for in-house tests.)

Across all languages tested, FSI's assessment of language proficiency relies primarily on in-person tests of speaking and reading, which have evolved from an approach first developed by the agency in the 1950s. Test scores are reported on a five-point scale (1 to 5) and defined using skill-level descriptions.[1] These descriptions were developed by the Interagency Language Roundtable (ILR), a group that coordinates second language training, acquisition, and testing approaches across the U.S. government. At the time of this report, the skill-level descriptions in the ILR framework were being revised.

The typical goal for language training is for Foreign Service officers to score at ILR level 3 in both speaking and reading (referred to as "3/3"), with this level of language proficiency intended to enable the kinds of job-related tasks the officers will encounter. Box 2-2 provides the ILR descriptions for level 3 reading and speaking, which are the focus of the FSI assessment. There are similar descriptions for listening, writing, translation, interpretation, and intercultural communication.

[1] The full skill-level descriptions for the ILR scale include a 0-level for no proficiency, "plus" levels for levels 0–4, and examples to elaborate the descriptions.

BOX 2-2
ILR Skill-Level 3 Descriptions for Reading and Speaking

Reading 3 (General Professional Proficiency): Able to read within a normal range of speed and with almost complete comprehension a variety of authentic prose material on unfamiliar subjects. Reading ability is not dependent on subject-matter knowledge, although it is not expected that the individual can comprehend thoroughly subject matter which is highly dependent on cultural knowledge or which is outside his/her general experience and not accompanied by explanation. Text types include news stories similar to wire service reports or international news items in major periodicals, routine correspondence, general reports, and technical material in his/her professional field; all of these may include hypothesis, argumentation and supported opinions. Misreading rare. Almost always able to interpret material correctly, relate ideas and "read between the lines," (that is, understand the writers' implicit intents in text of the above types). Can get the gist of more sophisticated texts, but may be unable to detect or understand subtlety and nuance. Rarely has to pause over or reread general vocabulary. However, may experience some difficulty with unusually complex structure and low frequency idioms.

Speaking 3 (General Professional Proficiency): Able to speak the language with sufficient structural accuracy and vocabulary to participate effectively in most formal and informal conversations in practical, social and professional topics. Nevertheless, the individual's limitations generally restrict the professional contexts of language use to matters of shared knowledge and/or international convention. Discourse is cohesive. The individual uses the language acceptably, but with some noticeable imperfections; yet, errors virtually never interfere with understanding and rarely disturb the native speaker. The individual can effectively combine structure and vocabulary to convey his/her meaning accurately. The individual speaks readily and fills pauses suitably. In face-to-face conversation with natives speaking the standard dialect at a normal rate of speech, comprehension is quite complete. Although cultural references, proverbs and the implications of nuances and idiom may not be fully understood, the individual can easily repair the conversation. Pronunciation may be obviously foreign. Individual sounds are accurate: but stress, intonation and pitch control may be faulty.

SOURCE: Excerpted from https://www.govtilr.org/Skills/ILRscale2.htm and https://www.govtilr.org/Skills/ILRscale4.htm#3.

Test Uses and Decisions

Scores on the FSI test are used to make many types of decisions about test takers. For example, many job postings to other countries are contingent on test scores. Foreign Service or other government agency personnel who go through the language training program before they leave to take up a posting in another country typically take FSI's language assessment at the end of the language training period. However, only about half of the tests administered each year are directly related to training. Non-training-related tests are taken by officers who want to add or update a score for retention or for promotion purposes or to have a current score on file so that they may apply for a posting in a country where that language is spoken.

The State Department provides requirements and incentives for personnel in language-designated positions to achieve a level 3 in both speaking and reading. Job assignments typically are contingent on achieving those target scores, although employees who do not reach the target ILR level of proficiency can take up their assignments while continuing to work to achieve the required proficiency.

In addition to the language score requirements associated with specific postings, employees receive incentive pay for their demonstrated proficiency in certain priority languages: 5 percent of salary for a 2/2 rating, 10 percent for a 3/3 rating, and 15 percent for a 4/4 rating. For all Foreign Service officers, scores below 4/4 expire after 5 years.

Components of the FSI Test

In all languages, the current FSI language proficiency assessment consists of a speaking test and a reading test. Listening is not tested separately but is incorporated in the speaking test. Although there are some variations, what follows is a general description of the FSI assessment.

Speaking Test

The speaking test has three parts:

1. **Social conversation.** The test taker introduces him or herself and discusses with the testing team topics such as daily life situations, and if proficiency allows, more complex topics, such as social, political, and current events.
2. **Work-related statement and exchange.** The test taker selects a general topic from a set of topics that are loosely aligned with the Foreign Service career tracks, such as consular affairs, diplomatic security, environment/science/technology/health care, international

development, management, political/military affairs, or public diplomacy. The test taker has 5 minutes to prepare an introductory statement on the topic. After the introductory statement, the test taker engages with the tester in the exchange part of the conversation.

3. **Interview.** This component of the test is an information gathering and reporting exercise. The test taker selects a topic from a category that aligns with the Foreign Service career tracks. Without preparation, the test taker begins interviewing the tester on that topic, in the language that is being tested. The test taker asks questions and listens to the responses until he or she feels that enough information has been collected. The test taker reports, in English, what was said immediately after the tester's response to each question.

Two aspects of the speaking test were changed in 2015. The social conversation now includes a gradual warm-up aimed at putting the test taker at ease, and a longer presentation task was replaced with a work-related exchange focusing more on an interactive dialogue.

Reading Test

The reading test consists of two tasks:

1. **Reading for gist.** This component is a carefully timed diagnostic test during which the testing team estimates the test-taker's working level in reading. The test taker is given six paragraphs of varying difficulty, with 6 minutes to identify the subject matter and the general meaning of as many passages as possible. Test takers are instructed that the task is like reading the newspaper—skimming and scanning documents for information.
2. **Reading in depth.** The outcome of reading for gist determines the level of difficulty of the text for the reading in depth portion of the test. Here, the test taker reads two to three longer articles in the target language and then reports, in English, on the main ideas, the supporting details, and information that generally explains the meaning of the text. The test taker is given 12 minutes to read each text. The objective is not to provide a direct translation of the text but instead for the test taker to use his or her own words to report as much information as possible from the text.

In 2018, the preparation time for the reading in-depth task was extended from 7 to 12 minutes to reassure test takers that they should focus on comprehension and not speed of reading.

Test Administration and Scoring

FSI strives for its test administration and scoring procedures to be consistent across all languages tested. Most FSI tests are conducted in person, by digital video conference, or by speakerphone. Although the preferred mode is in person, video conferencing has increased in recent years and is now used for about 20 percent of test takers, and testing by speakerphone is around 10 percent.

The speaking and reading tests each last about 1 hour. Test takers can start with either portion, and the two portions can be separated by an optional 5-minute break.

The test is administered to individual test takers by a tester and an examiner. The tester is the rater who interacts with the test taker in the language of the test. The examiner interacts with the test taker in English to administer the test, provide instructions, and monitor the timing of each task. FSI's goal is for testers and examiners to be unfamiliar with test takers. This goal is relatively easily accomplished in high-volume languages, for which the language school has full-time testers and multiple instructors. However, for languages with fewer learners and thus fewer instructors, the tester may have also been the test-taker's language teacher in the early phases of language training.

In contrast with other agencies that use the ILR framework for language proficiency testing, FSI does not align specific reading texts with individual ILR levels. Based on the FSI's belief that it is possible to show a range of proficiency when reading a specific text, there are three general categories of FSI reading texts that roughly correspond to the proficiency ranges of ILR levels 1 to 2 (A-level texts), 2-plus to 3 (B-level texts), and 3-plus to 5 (C-level texts). FSI's testing protocol is adaptive in that it involves an initial determination of the working level of the test taker and includes the flexibility to move up or down from that initial level.

The FSI scoring approach also differs from other agencies. Scoring is an interactive, deliberative, and consensus-based procedure involving the tester and examiner (Hart-Gonzalez, 1994). An overall ILR proficiency-level rating is determined holistically, with the tester and examiner reaching an initial tentative consensus, based on their overall judgment of the performance. They then consider the test-taker's strengths and weaknesses related to five factors: comprehension, ability to organize thoughts, grammar, vocabulary, and fluency. As part of this consideration, the tester and the examiner separately estimate quantitative values for the five factors, which are added together to create an "index" score on a 120-point scale. This index score is used as a check on the holistic rating on the ILR scale and to confirm the consensus between the tester and the examiner, possibly leading to some adjustment in the consensus ILR scale score. Although listening is not considered explicitly or reported separately, listening skills are obviously required to perform well on the speaking test and are reflected in the comprehension factor. The scoring sequence—from initial holistic rating to the five-factor derivation of an index score and then to comparison of the separate index scores with the initial holistic rating—is repeated two times, once for the speaking test and once for the reading test.

If test takers are dissatisfied with their test results, they can ask for their scores to be reviewed within 30 days of their test. They can generally retake the test after 6 months.

3

Language Constructs and Language Use

As illustrated in the committee's guiding framework (see Figure 1-1, in Chapter 1), the design for a high-stakes language assessment for use in a professional setting starts from an understanding of the nature of language and its use, the broader sociocultural and institutional contexts for the assessment, and the specific language use in the domain that will be targeted for the assessment. This chapter discusses some of the key concepts and techniques that inform these understandings.

LANGUAGE CONSTRUCTS

The knowledge, skills, abilities, and other characteristics that are the focus of an assessment are described in terms of a "construct." A construct is an abstract, theoretical concept, such as "knowledge" or "proficiency," that has to be explicitly described and specified in test design. This definition usually comes from a mix of theory, research, and experience.

Construct definition plays a central role in a principled approach to assessment design and use. The goal of defining the construct is to provide a basis not only for the development of a given assessment but also for the interpretation and use of its results. For FSI, the construct will relate to descriptions of the language proficiency of Foreign Service officers who need to use a given language at a foreign post.

Conceptualizations of language and language proficiency become more nuanced over time, so every testing program needs to periodically revisit its construct definitions. Since the 1960s, approaches to construct definition have evolved to reflect broadened conceptions of language and language use. They also reflect ongoing refinements in language assessment theory, advances in theories of learning and knowing, especially with respect to context and the social dimension of language, and the changing nature of language assessment in light of advances in technology (Bachman, 2007; Purpura, 2016). These refinements have had important consequences for operationalizing the construct of language proficiency and conceptualizing and justifying validity claims, and are, to varying degrees, reflected in current language assessments.

To address FSI's desire to keep pace with developments in language assessment, this section summarizes four key approaches to defining language proficiency and their implications for the design and operationalization of test constructs and for the meaningful interpretation of performance on a test: trait-based, interactionist, meaning-oriented, and task-based. This summary illustrates the expansion of the construct of language proficiency over time, but the committee is not suggesting that all assessments should use the broadest measure possible. Rather, we call attention to the many different factors that can be considered in an assessment, depending on its intended goals and uses, and highlight the importance of being explicit about these factors and their contribution to performance. Such careful attention to the intended construct will allow for an accurate representation of a scored performance and its meaningful interpretation.

Trait-Based Approach

Probably the oldest and most common approach to defining the construct of language proficiency is to specify in a theoretical model how the trait of language proficiency is represented in a test-taker's mind. This is done by identifying the knowledge components—such as grammatical knowledge—that underlie a test-taker's proficiency and then designing tasks that measure those components ("traits"). Lado (1961) used this approach to conceptualize language proficiency as the ability to use language elements or forms (grammatical forms, lexical meanings) to understand and express meanings through listening, reading, speaking, and writing. Carroll (1961) expanded this conception to include not only *how* individuals communicate but also *what* they communicate in a complete utterance.

Knowledge of the mapping between form and meaning is still a critical component of language use (VanPatten et al., 2004), and it is the basis for grammatical assessment in tests designed to measure grammatical ability (e.g., the Oxford Online Placement Exam[1]). It has also been a central feature of scoring rubrics (scoring instructions and criteria) of language pro-

[1] See https://www.canada.ca/content/dam/ircc/migration/ircc/english/pdf/pub/language-benchmarks.pdf.

ficiency that have an independent language use scale (e.g., the TOEFL iBT test[2]); rubrics that have grammatical-level performance descriptors (such as in the skill-level descriptions of the Interagency Language Roundtable [ILR], used by FSI and discussed in Chapter 2); and approaches to the automatic scoring of speaking and writing (Purpura, 2004, 2016). Knowledge of this mapping is also reflected in the widely used "CAF" measures, which incorporate measures of three related but separable traits: complexity, accuracy, and fluency in speaking or writing (Housen et al., 2012). However, this conceptualization fails to resolve the potential vagueness and ambiguity of meaning often found in language.

Notable expansions of the language proficiency trait beyond grammatical rules and vocabulary include *communicative competence* (Canale, 1983; Canale and Swain, 1980) and *communicative language ability* (Bachman, 1990; Bachman and Palmer, 1996, 2010), which incorporate additional components to the language use model, such as knowledge of how to use language to achieve a functional goal or how to use language appropriately in social contexts with a diverse range of interlocutors. Bachman's communicative language ability model specifies grammatical knowledge, textual knowledge, functional knowledge, and sociolinguistic knowledge. It has been used, for example, to guide the development of the Canadian Language Benchmarks Standards for adults learning English as a second language.[3] Alongside language knowledge, this model also specifies the role that strategic processing plays in the ability to complete language-related tasks, which underlies the examinee's ability to consider context, content, language, and dispositions while generating responses during a test, all considerations in the skill-level descriptions used by FSI.

Interactionist Approach

Despite its strengths, the trait-based approach does not fully specify how language ability is affected by the context of language use. Context is a key determinant of language use, as can be seen in the Foreign Service context by the contrast between informally communicating with host nationals in a coffee shop and interacting with high-ranking government officials in a policy discussion. The *interactionist approach* (Chapelle, 1998) to construct definition addresses this omission by highlighting the role that the features of context, in addition to language knowledge and strategic processing, play

[2] TOEFL is the Test of English as a Foreign Language; TOEFL iBT measures one's ability to use and understand English at the university level. See https://www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf.

[3] See https://www.canada.ca/content/dam/ircc/migration/ircc/english/pdf/pub/language-benchmarks.pdf.

in language proficiency. With this approach, according to Chalhoub-Deville (2003), language proficiency is seen as an "ability-in-individual-in-context."

Recognizing that the nature of language knowledge differs from one domain of use to another, Douglas (2000) proposed the *language for specific purposes framework*. In this framework, language ability reflects the interaction among language knowledge, specific-purpose background knowledge, the contextual features of specific-purpose language use, and the ability to put all these components together simultaneously through strategic processing. In this framework, all the components should be specified and accounted for in test development and interpretation. The FSI test is a form of language test for specific purposes in that many of the tasks reflect characteristics of the Foreign Service context and require test takers to engage language and content knowledge specific to Foreign Service work.

Meaning-Oriented Approach

Extending the interactionist approach, a *meaning-oriented approach* to the construct definition of language proficiency added "the ability to effectively express, understand, dynamically co-construct, negotiate, and repair variegated *meanings* in a wide range of language contexts" (Purpura, 2017, p. 1). In other words, this approach underscores the role of meaning and the communication of literal and contextually constructed meanings in the conceptualization of language proficiency.

The meaning-oriented conceptualization of language proficiency provides a detailed depiction of the knowledge components underlying language use. It suggests that, depending on the assessment task characteristics, contextualized performance could be observed and scored for (1) grammatical accuracy, complexity, fluency, or range; (2) content accuracy or topical meaningfulness; (3) functional achievement or task fulfillment; and (4) pragmatic appropriateness (e.g., formal register) (for further details, see Purpura and Dakin, 2020). This model is also useful for assessments that seek to use independent and integrated skills rubrics to hold test takers responsible for topical information presented in the assessment itself (as in the TOEFL iBT test noted above). It has also been useful for conceptualizing and scoring the ability to understand and convey nuanced pragmatic meanings implied by context (e.g., sarcasm).

Task-Based Approach

The approaches discussed so far attribute performance consistencies to expressions of the knowledge, skills, abilities, and other characteristics that test takers have and can apply during language use. All of these play a role in a test-taker's ability to perform tasks on the current FSI assessment, and some of them are incorporated into the assessment's scoring rubric. In contrast, a different approach emerged in the 1990s that focused on test-takers' ability to successfully complete tasks that approximate real-world instances of communicative language use designed for specific purposes in given contexts (Brindley, 1994). As this approach mostly uses "task performance," not "language knowledge or communicative language ability," as the unit of analysis (Brown et al., 2002; Long and Norris, 2000; Norris et al., 2002), it is called a *task-based approach* to construct definition.[4]

The task-based approach seeks to create assessment conditions that approximate real-life contexts in which the tasks "replicate typical task procedures, content, situations, interlocutors, and other factors, in order to provide trustworthy indications of the extent to which learners can handle the real-world requirements of task performance" (Norris, 2016, p. 236). Norris et al. (2002) implemented this approach in a rating scale designed to evaluate test-takers' success in responding in writing to a voicemail request from a boss to make a hotel reservation. In this example, the rating scale ranges from "inadequate" to "adept" performance. At the lower end of the scale, inadequate responses involve the choice of an incorrect hotel, a confusing message about the reservation, or a stylistically inappropriate message. At the higher end, adept responses involve a correct choice for the hotel, a clear justification for the decision, and a stylistically appropriate message.

The task-based approach has contributed to the scope of language assessment by highlighting the importance of functional language use based on task fulfilment. This approach corresponds with the notion of task accomplishment as the desired standard or outcome; it is reflected in the performance descriptors of many assessment frameworks that focus on observation of the outcome. For example, did the test taker succeed in describing the advantages and disadvantages of the U.S. educational system to hypothetical host-country nationals during the test? A "pure" or "strong" task-based approach may consider only the task outcome and not the language the test taker used (Clark, 1972; Long, 2015; McNamara, 1996); other versions consider task fulfillment alongside knowledge components of language use as part of a task-based construct.

[4] A separate approach to using task-based assessment valued "tasks" for their potential to trigger cognitive processes related to language rather than because of their potential to provide estimates of real-world language use (see, e.g., Skehan, 1998, 2003; Robinson, 2001).

CURRENT UNDERSTANDINGS OF LANGUAGE AND LANGUAGE USE: IMPLICATIONS FOR DEFINING THE CONSTRUCT OF LANGUAGE PROFICIENCY

As the above discussion illustrates, language is no longer viewed from a uniquely cognitive perspective as a set of discrete linguistic forms and skills that need to be mastered. Instead the field is moving toward a more sociocultural perspective, in which language is viewed as a complex system of communication that is often constructed through use. Indeed, a recent analysis of 42 journals in applied linguistics (Lei and Liu, 2018) found that such topics as traditional phonological and grammatical issues have decreased significantly since 2005. Instead, Lei and Liu (2019, p. 557) note:

> [T]he most popular topics now include the impacts of socioeconomic class, ideology, and globalization on language use/learning and identity in various local contexts, the development and use of English as a Lingua Franca, the practice and effects of multilingualism, and corpus-based investigation of field-specific discourse and literacy practices and variations.

The sociocultural perspective considers language as "a resource for participation in the kinds of activities our everyday lives comprise" (Zuengler and Miller, 2006, p. 37). This perspective highlights the multifaceted nature of language and its use in the real world. Important dimensions of the sociocultural perspective include the value of multiple varieties of any given language use, the increasingly multilingual nature of language use, and the recognition that communication is multimodal.

The idea of the value of multiple varieties[5] of any given language reflects an important shift in assessment away from the notion of a "native speaker" as the gold standard of language proficiency (Davies, 2003; Dewaele, 2018). For example, in the context of learners of English, instead of viewing language learners as having a *deficit* linguistic variety (Kachru, 1996), some applied linguists argue that English belongs to anyone who uses it (Seidlhofer, 2009). In this view, international or World English(es) are accepted as complete and whole linguistic systems of communication that have no bearing on when, how, or by whom the language was learned (Jenkins, 2006).

[5] In sociolinguistics, "variety" or "dialect" is a general term for any distinctive form of a language. Wolfram et al. (1999, p. 3) defined "language variety" (which at the time was used synonymously with "dialect") as "a social or geographic form of language distinguished by the specific pronunciation, vocabulary, and grammar of its speakers." As a geographic example, on a narrow scale, a New York variety of English is different from a Texas variety of English. On a broader scale, an American variety of English is different from a British variety of English. Social examples include varieties used by a socioeconomic class, a profession, an age group, or any other social group (Nordquist, 2020).

The increasingly multilingual nature of language use reflects the fact that "there are almost 7,000 languages in the world and about 200 independent countries" (Cenoz, 2013, p. 3), suggesting that multiple languages are likely used in any given country and that many individuals are likely multilingual. Moreover, multilingual individuals often use multiple languages simultaneously in a conversation or in writing, drawing on all their linguistic repertoire in constructing meaning (translanguaging). Globalization, immigration, and new technologies have contributed to the growing importance of multilingualism in modern society. Given this reality, there have been calls for language assessments to reflect the multilingual nature of communication in various contexts of target language use and the language abilities of multilingual users (Gorter and Cenoz, 2017; Schissel et al., 2019; Shohamy, 2011).

It is now recognized that communication is multimodal, and language use is just one resource for achieving meaning. A common view among applied linguists and language educators is that language is the primary means for communicating meaning. This view continues to be challenged, however, and replaced by the idea that meaning is communicated through both linguistic and nonlinguistic modes (e.g., images, gestures, three-dimensional models) that are socially and culturally shaped (Kress, 2010). This expanded view emphasizes the relationships between and among linguistic modes (e.g., comparisons of listening, speaking, reading, and writing) to accomplish communicative goals. It also includes attention to nonlinguistic modes because the potential for conveying meaning is increased when they are used with linguistic modes (Lemke, 2002).

These contemporary understandings of language use—involving not just varieties of a language but also multiple languages and modalities—have implications for assessment and are already being reflected in some assessments. For example, the TOEFL iBT now uses varieties of English from North America, the United Kingdom, New Zealand, and Australia as test inputs. Some language testing researchers also are beginning to design language tests that include translanguaging components in the test input and that allow for translanguaging in the response, thus "enabling test takers to draw on their entire repertoires as multilingual persons, and more authentically representing and valuing the translanguaged reality of current workplace language practice" (Baker and Hope, 2019, p. 421). Finally, the idea of multimodal communication is reflected in the increasing use of integrated tasks in language assessment.

These broader understandings of language use also have prompted calls for broadening language constructs in assessment. For example, the Modern Language Association (MLA) Ad Hoc Committee on Foreign Language (2007) has called for an emphasis on "translingual and transcultural competence," which it defines as the "ability to operate between languages."

Focusing specifically on English, Hu (2018, p. 80) has proposed the construct of "communicative effectiveness" that would take into account, among other things, "the necessity of an empathetic openness to different varieties of English, the relevance of various dimensions of understanding and the crucial role of strategic competence in negotiating communication online."

The current FSI test already embraces multilingual perspectives to some degree. In two sections of the test, test takers are required to use two languages: in the interview section of the speaking test, they interview the tester in the tested language and report what they learn to the examiner in English; in the reading in depth section of the reading test, they read a passage on a specialized topic in the target language and then summarize it orally in English. These tasks likely also occur in a similarly multilingual way in the daily work of Foreign Service officers.[6]

LANGUAGE USE IN PROFESSIONAL SETTINGS

Moving from considering language use in a broad sense to its use in a specific work-related or professional context—in FSI's case, the use of language in Foreign Service tasks—raises a separate set of assessment issues. These issues relate to the use of the test scores for high-stakes employment-related decisions and the procedures for determining the scope of tasks covered on the test.

Scores on the FSI test are used to make many types of personnel decisions about job placement, promotion, retention, and pay. Principled approaches to assessment and professional standards for assessment and evaluation suggest that assessments that are used to support personnel decisions in professional settings should be grounded in an understanding of real-world job requirements to be valid and legally defensible. The U.S. government's generally accepted guidelines on decisions involving selection, promotion, retention, pay, and other related career decisions emphasize the need to demonstrate close approximation between the content of a test and the observable behaviors or products of a job.[7] Moreover, validity is enhanced when an assessment aligns to the work context (Sackett et al., 2016). By extension, the content of language tests in professional settings should be relevant to the types of decisions that test results are used to make.

[6] In the context of the Common European Framework for Reference, such multilingual tasks are understood to require "mediation" between two languages (Council of Europe, 2018).

[7] See http://uniformguidelines.com/uniformguidelines.html#67.

Job Analysis and Assessment Development

Job analysis is one way to connect language use in a professional setting to the content and format of a language test. Broadly speaking, understanding the content of a job, set of jobs, or an occupation involves standard work or job analysis techniques. Job analysts use these techniques to identify tasks or work behaviors and the knowledge, skills, abilities, and other characteristics of workers that underlie performance on these tasks (Brannick et al., 2017). Knowledge, skills, abilities, and other characteristics refer to the characteristics that workers need for either performing the tasks at hand or demonstrating the human behaviors described in the job analysis. These characteristics are generally considered to be constructs—as defined in psychological and educational measurement—that predict job performance. Job analysis can also document the physical and physiological context in which the work is performed, such as stressful, ever-changing, or extreme contexts.

Specifying the critical tasks and identifying the underlying knowledge, skills, abilities, and other characteristics that enable performance of the tasks are important to any kind of test development. Linking test content to tasks helps to establish content validity evidence, while linking test content to important worker characteristics helps to determine the specific constructs that a given test needs to measure. In addition, job analysis makes it possible to build a test plan or "blueprint" showing the relative importance or weight of different topics that correspond to tasks and knowledge, skills, abilities, and other characteristics, which helps ensure that the job domain has been sampled adequately (Brannick et al., 2017). Job analysis can also illuminate how real-world job aids are used—such as online translation programs—and to understand how a job is changing and could require future changes to an assessment.

It is important to note that not all knowledge, skills, abilities, and other characteristics that are identified in job analysis need to be tested, depending on the employee population and the types of training that may be provided. However, job analysis can identify the set of characteristics that employees need and that should be considered for testing. In the FSI context, the critical characteristics to consider for language proficiency testing will involve tasks that are carried out using a foreign language.

The techniques for conducting job analysis are too voluminous to review here. However, a few notable methods that could be used in a foreign language assessment context to infer specific language demands from known job demands include

- evidence-centered design (Mislevy et al., 1999a, 2003), a structured assessment development process to ensure that the evidence gathered from the assessment tasks corresponds to the underlying constructs that the assessment purports to address—in this case, language use in the professional setting;
- ethnographic approaches, which investigate the nature, type, and quality of workplace language through methodologies that illuminate social processes for individuals in workplace contexts (Newton and Kusmierczyk, 2011);
- critical-incidents techniques to generate lists of examples of especially good and poor job performance behaviors (or "incidents") and identify observable behaviors that may lead to overall success or failure in a position (Gatewood et al., 2015); and
- cognitive task analysis, which uncovers the knowledge structures that people use to perform tasks and helps elucidate contextual factors that affect performance (Mislevy et al., 1999b).

Regardless of the technique, one key design decision involves the level of specificity of analysis. Jobs can be studied at various levels, from specific positions in one organization to occupations that describe the entire economy. Similarly, knowledge, skills, abilities, and other characteristics can be described narrowly or broadly: for example, speaking skill could be described for a specific role, such as customer service, or broadly, across all possible contexts. Ultimately, job analysts and assessment developers must specify the domain of use and the degree of generalization that is assumed across tasks. Box 3-1 provides an example of language use in the Foreign Service context, illustrating some of the specific features of the domain, which would need to be clarified in a job analysis.

Job Analysis and Language Assessment for Professional Purposes

Target language use analysis is an examination of tasks and knowledge, skills, abilities, and other characteristics that are relevant to the development of language tests (Bachman and Palmer, 1996, 2010). Target language use analysis uses a framework of task characteristics to identify critical features of occupational, academic, or other language use in real-world contexts (Bachman and Palmer, 2010). Test developers can use this framework to describe characteristics of the language use setting, characteristics of the inputs and expected responses, and relationships between inputs and expected responses. Test developers can use these same characteristics to specify language use tasks for assessment, maximizing approximations between the actual context and the assessment tasks. Existing research

BOX 3-1
Example of Language Use by a Foreign Service Officer

Upon arriving in the heavily patrolled Zone, I soon realized that I was the only English speaker for miles around. In discussions with the Comandante, he frowned when I told him that the "Embajador" would be arriving by helicopter.

"Is he coming in a green military helicopter?" he asked in Spanish.

"Oh no," I assured him. "He would be coming in a yellow civilian helicopter."

"That is good," he said, telling me something about helicopters in rapid Spanish. Depending upon the verb-ending, he was either saying: "We *used* to shoot at helicopters," "We *do* shoot at helicopters," or "We're *going* to shoot at helicopters."

Cursing myself for not having paid better attention in my Spanish classes at the Foreign Service Institute, I tried to clarify, saying "But tomorrow we are *not* shooting at helicopters."

"No," he laughed. "No problem."

SOURCE: ACT, Inc. (2012, p. 3).

that describes aspects of the target language use domain relevant to the FSI's Foreign Service context can serve as a useful resource (e.g., Friedrich, 2016). A test blueprint (for an example, see So et al., 2015) could be built based on information combined from job analyses and target language use analyses. In a work setting, developers can identify subskills and stimulus situations directly from job analysis—using tasks and knowledge, skills, abilities, and other characteristics—and weight these elements according to their importance to overall job functioning, creating linkages that support validity argumentation.

A recent approach for conceptualizing professional communication identified four different varieties of language ("codes of relevance") that can inform the development of language assessment for professional purposes (Knoch and Macqueen, 2020):

- Intraprofessional language is used by a small number of people with shared professional knowledge (e.g., doctors speaking to each other in "medicalese"). Language use is practically inseparable from content knowledge.

- Interprofessional language involves interactions among individuals with some shared professional knowledge (e.g., a doctor interacting with a nurse or social worker in "cross-disciplinary medicalese").
- Workplace community language involves interactions between those with professional knowledge and lay people (e.g., a doctor communicating with a patient).
- Language varieties used in the broader social context include all language varieties and minority languages in the jurisdiction, as well as whatever patterns govern their use and combination, which can illuminate where miscommunications occur in the workplace and how they can be reduced.

Sampling from these different language varieties to develop a language assessment for professional purposes involves careful analysis of the professional context (the target language use domain) and the purpose of the assessment (Knoch and Macqueen, 2020).

In terms of sampling the job domain, Foreign Service jobs vary across several dimensions, such as career tracks, specialist or generalist, and differences in language requirements. Every job analysis needs to consider differences in job requirements across these dimensions and how these differences may be reflected in the test specifications. Moreover, it is worth noting that FSI uses the ILR framework to designate job language requirements for language-designated positions. Thus, deliberations about the role of the ILR framework now and in the future should consider that the ILR describes not only worker-related requirements (skills) but also work or job requirements.

OTHER CONSIDERATIONS IN DEFINING THE CONSTRUCT OF LANGUAGE PROFICIENCY

Test scores reflect an examinee's proficiency with regard to the construct that is explicitly assessed, as well as other factors that are not intended to be explicitly measured (Bachman, 1990; Turner and Purpura, 2016). For example, the current FSI test contains a speaking component, which is designed to determine whether test takers have sufficient proficiency in a language to gather information from an interlocutor, retain that information, and then report back in English to another interlocutor. Although oral language proficiency represents the *proficiency dimension* of the assessment and is the explicit object of measurement, performance on the test can be influenced by other factors, such as the test-taker's memory, processing skills, affective dispositions, and task engagement. Although it might appear that language testers are only measuring the construct of language proficiency because they score responses related to the proficiency dimension, these other factors are also involved, and they often moderate performance. These factors (called *performance moderators* by O'Reilly and Sabatini, 2013) can enhance or detract from the measurement of proficiency. Purpura and Turner (2018) elaborate on five types of performance moderators:

1. The contextual dimension addresses the social, cultural, institutional, political, or economic characteristics of the assessment context and the extent to which these characteristics might impact performance.
2. The sociocognitive dimension includes the extent to which test takers have the mental capacity to process, retain, and retrieve information, and the capacity to execute those steps with automaticity. This dimension is also invoked in assessments where test takers receive feedback or assistance that they are expected to process in order to improve their performance.
3. The instructional dimension reflects the need for a test taker to process new information.
4. The social-interactional dimension reflects the extent to which the test taker needs to manage interaction, such as turn-taking (Levinson, 1983).
5. The affective dimension addresses the effect of the test-taker's engagement, effort, anxiety, confidence, and persistence on test performance.

Traditional assessment design frameworks often focus on the context, elicitation, and proficiency dimensions of assessments. However, many fail to explicitly address these other factors in the design stage, even though they can affect performance. Whether or not these moderators are defined as part of the test construct and are explicitly measured, their implications for test design, development, and validation need to be considered.

4

Possible Approaches for Assessing Language Proficiency in the FSI Context

This chapter turns to possible changes to the design of FSI's assessment of foreign language proficiency, including consideration of tasks that could be used in the assessment, the performances they can elicit, and the scores that result from the evaluation of those performances. It builds on the two previous chapters: the FSI testing context (Chapter 2) and the nature of language and language use (Chapter 3).

OVERVIEW

Scope of Committee's Work

The chapter's discussion of possible changes to the current FSI test is guided by the description of the considerations in language assessment development in Figure 1-1 (in Chapter 1). The presentation of possible changes is based on the assumption that an analysis of the test's key validity claims has been carried out, guided by a principled approach. Such an analysis would look at the relationships among the components of the assessment, the way the resulting test scores are interpreted and used, the target language use domain, the sociocultural and institutional contexts of the test, and the current understanding of language proficiency (see discussion in Chapter 6).

The chapter does not present a comprehensive description of all possible methods for assessing language proficiency. For surveys of the literature, the reader can consult any number of authoritative textbooks and handbooks, including the following: for overall treatments of language assessment, Bachman and Palmer (2010), Fulcher and Davidson (2012), Green (2014), and Kunnan (2018); for assessing speaking, Luoma (2004), Taylor (2011), and Taylor and Falvery (2007); for assessing writing, Cushing-Weigle (2004), Plakans (2014), Taylor and Falvery (2007), and Weigle (2002); for assessing listening, Buck (2001) and Ockey and Wagner (2018b); for assessing reading, Alderson (2000); for assessing grammar, Purpura (2004); for assessing vocabulary, Read (2000); for assessing integrated tasks, Cumming (2013), Gebril and Plakans (2014), and Knoch and Sitajalabhorn (2013); and for assessing language for professional purposes, Knoch and Macqueen (2020).

Rather than describing all possible approaches to assessment, the committee selected the changes to discuss on the basis of its knowledge of the research in language assessment and its understanding of FSI's context, target language use domain, and current test. Each of these possible changes is keyed to possible goals for improvement that they might address, while also noting their implications for other important considerations. The committee does not suggest ways that the different possible changes might be integrated into the current test. Some of the changes are minor alterations to current testing procedures; others involve wholly new tasks or testing approaches that might complement or substitute for parts of the current test.

The discussion of each possible approach provides examples of testing programs that have implemented the approach and relevant research references, as they are available. However, it is important to note that the field of language testing often proceeds by developing approaches first and then carrying out research as experience accumulates related to the innovations. As a result, the different possible approaches discussed below reflect a range of available examples and supporting research.

The first possible change we discuss below is multiple measures, which can be understood as a meta-strategy to combine different tests with complementary strengths to produce a more complete and comprehensive assessment of an individual than is available from any single test. The possible use of multiple measures is the first change discussed because many of the other changes could be carried out alongside the tasks or approaches of the current test. Whether to think of the possible changes as complements or substitutes to the current test is one of the choices FSI would need to consider during a program of research and development.

All the other possible changes were chosen in response to two plausible goals for improvement that might emerge from an evaluation of the complete context of the FSI test: (1) broadening the construct of the test or the test's coverage of that construct, and (2) increasing the reliability of the test scores and the fairness of their use. It is important to note that any changes to the test will have other effects beyond these two goals. In the discussion of possible changes, the committee considers two particular ones: effects of the test change on instruction and practical effects related to the cost and feasibility of the testing program. In another context, these two effects could themselves be considered the primary goals for the improvement of a testing program; however, in the context of FSI's request, the committee has taken potential instructional and practical effects as important additional considerations, not as the primary goals for improvement likely to emerge from an evaluation of the current test. Before considering the specific possible changes to the current test, we elaborate on these two goals and effects.

Goals and Effects of Possible Changes

As discussed in Chapter 3, the construct measured by a test and its alignment with the target language use domain are critical to the validity of the inferences about a test-taker's language proficiency in that domain. A consideration of the range of different language constructs stemming from job and target language use analyses related to the specific language needs of Foreign Service officers could suggest aspects of language proficiency that are important in Foreign Service work but that are not reflected, or perhaps not sufficiently reflected, in the current test. Listening proficiency is an example of an aspect of language proficiency that is perhaps not sufficiently reflected on the current test, and writing proficiency is an example of an aspect of language proficiency that is not reflected at all. In addition to these clear examples, the committee's consideration of the Foreign Service context and the FSI test indicated some other possible improvements related to the assessed construct, depending on FSI's goals. These other possible improvements include assessing interactional competency in more natural social exchanges, assessing listening proficiency across a typical range of speech, and assessing the ability to use language in tasks that are similar to those required on the job. Such improvements might strengthen the assessment of aspects of the language proficiency construct that are already partly reflected on the current test.

With respect to the first goal of broadening the construct measured by the test or the test's coverage of that construct, several possible changes are discussed below: scoring for listening comprehension on the speaking test, adding writing as a response mode for some reading or listening tasks, adding paired or group oral tests, including listening tasks with a range of language varieties and unscripted texts, incorporating language supports, adding a scenario-based assessment, and incorporating portfolios of work samples.

The second goal, increasing the reliability of the test scores and fairness of their interpretation and use, is partly reflected in FSI's request to the committee and is a plausible goal that might emerge from an internal evaluation of the current testing program. Some considerations that might lead to such a goal are discussed in Chapter 6, relating to such factors as the criteria used in the scoring process and the consequences of the decisions based on the test. High levels of variability in the test scores could give rise to concerns among stakeholders about reliability, and differences across individuals that reflect factors other than language proficiency may suggest concerns about fairness and possible bias.[1] General approaches for increasing fairness and reliability in a testing program involve standardizing aspects of the test and its administration, scoring, and use; being transparent about the test and its administration, scoring, and use; and using multiple testing tasks, administrators, formats, and occasions.

It is important to note that there can be some tension between these two goals to the extent that the richness of more natural language interactions can be difficult to standardize. Obviously, it is not productive to standardize a language proficiency test in the service of the second goal in ways that prevent the test from assessing the aspects of language proficiency that are important in the target language use domain.

With respect to the second goal of increasing the fairness and reliability of test scores, the discussion below covers the following possible changes: adding short assessment tasks administered by computer, using automated assessment of speaking, providing transparent scoring criteria, using additional scorers, and providing more detailed score reports.

The structure of a test is often a powerful signal to its users about the competencies that are valued, which can have important effects on instruction. Within the field of language studies, these effects are sometimes referred to as washback. The effects can have both positive and negative aspects: positive when the test signals appropriate goals for instruction and negative when the test encourages narrow instructional goals that fall short

[1] One form of bias—“unconscious bias” or “implicit bias”—concerns evidence of unconscious preferences related to race, gender, and related characteristics (Greenwald et al., 1998). For example, participants might automatically associate men with science more than they do for women. One can easily imagine the possible problematic effects of such associations in a language testing context, such as unintended inferences on the part of interviewers, raters, test takers, or test designers. Despite concern over unconscious bias, reviews of hundreds of studies conducted over 20 years reveal two key findings: the correlation between implicit bias and discriminatory behavior appears weaker than previously thought, and there is little evidence that changes in implicit bias relate to actual changes in behavior (Forscher et al., 2016; Oswald et al., 2013). More research is needed to clarify these findings. Regardless of the degree to which unconscious or implicit bias affects real-world employment and testing conditions, the best practices in assuring test fairness, highlighted throughout this report, remain the best defense against such effects.

of the language proficiency needed to accomplish tasks in the real world. Although the committee did not consider possible changes specifically to provide positive effects on instruction, a number of the changes considered as possible ways to meet the primary goals would also likely have a positive effect on instruction, which are discussed below when this is the case.

Finally, any changes to the test will raise practical considerations, including effects related to cost and operational feasibility. As with instructional effects, the committee did not specifically consider possible changes to the test with a goal of decreasing its cost or maximizing the ease of its administration. However, the discussion below notes the potential practical implications for the changes suggested to meet the two primary goals.

Although cost is an obvious practical consideration, in the FSI context it is important to highlight a perhaps more fundamental consideration, which is the wide range in the number of test takers across languages: from over 1,000 annually for Spanish, to less than 10 annually for several dozen languages, such as Finnish, Mongolian, and Tagalog. For the many languages with few test takers, there are fewer speakers who can serve as testers and test content developers, fewer resources in the language to draw on, and fewer opportunities to try out any new testing techniques with test takers. For all these reasons—in addition to the more obvious one of the associated cost that would be involved—the hurdle for implementing possible changes to the test will be much higher for any possible change that involves developing new testing material for all of FSI's tested languages. To the extent that it is necessary to keep the structure of the test identical across all tested languages, the practical feasibility of any possible changes for the languages with few test takers could be an important constraint. This issue as to whether test structure must be held constant across all languages is discussed further in Chapter 7; in the discussion below we simply note the possible changes that may raise particularly high practical difficulties for those languages with few test takers.

Table 4-1 summarizes the possible changes to the FSI test that are discussed in the rest of this chapter, with the potential goals the committee explored that might motivate consideration of these changes and their additional instructional or practical considerations. As discussed above, the potential goals for change would need to emerge from an overall review of the current test and its context, using a principled approach to analyze how the current test could be strengthened to support the key claims about its validity. The possible changes are listed in the table in the order they are discussed in the rest of the chapter.

TABLE 4-1 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

TABLE 4-1 Continued

Possible Change	Potential Test Construct, Reliability and Fairness Considerations	Potential Instructional and Practical Considerations
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 • 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

USING MULTIPLE MEASURES

The fundamental idea behind the use of multiple measures is to make decisions on the basis of results from several different assessments. By using multiple measures, a testing program can expand coverage of the construct by combining information from different sources that assess different aspects of it. Reliance on multiple sources, such as several assessments that use different response modes, can help ameliorate the effects of any particular source of error. This can help increase overall reliability, generalizability, and fairness.

The value of using multiple measures in an assessment is strongly supported in the research on testing (Chester, 2005; Kane, 2006; Koretz and Hamilton, 2006; Messick, 1989). Several of the professional testing standards explicitly call for the use of multiple measures in testing programs that are used to support high-stakes decisions (e.g., American Educational Research Association et al., 2014; National Council on Measurement in Education, 1995; also see Chapter 6). These standards are reflected in current K–12 educational policy legislation, such as the Every Student Succeeds Act. Although there are some important measurement issues that need to be addressed when combining test scores (Douglas and Mislevy, 2010; In'nami and Koizumi, 2011), there are examples of practical implementations of decision systems using multiple measures (e.g., Barnett et al., 2018).

For FSI, additional measures might be added to complement the current speaking and reading tests in response to goals that are identified from a review of the test and its use. A number of the possible changes discussed below provide examples of additional measures that could be added to the current test to produce an overall testing program using multiple measures.

SCORING LISTENING ON THE SPEAKING TEST

The committee's statement of task specifically asks about the possibility of explicit assessments of listening comprehension that could be part of the FSI assessment. In reviewing the options, the committee noted that FSI could augment the scoring for the speaking part of the test to make more use of the information related to listening that it already provides. Specifically, the three tasks in the speaking test could be explicitly scored in relation to listening proficiency, with reference to the Interagency Language Roundtable (ILR) skill-level descriptions (see Chapter 3) for listening and the development of a set of listening-related scoring criteria (Van Moere, 2013). This approach might add some additional complexity to the scoring process, but it would not necessarily require much change in the tasks themselves.

The FSI speaking test is notable in using three different speaking tasks that each involve a variety of language-related skills. In this sense, the current test reflects the new research on integrated skills that recognizes that most language tasks in the real world require multiple skills (e.g., Cumming, 2013; Knoch and Sitajalabhorn, 2013; Plakans, 2009). According to this view, the best tasks to assess a target language use domain will be tasks that integrate multiple skills. For FSI, for example, a Foreign Service officer participating in a formal meeting might need to make an initial presentation and then respond to follow-up questions, a use of integrated skills echoed in the "work-related exchange" portion of the FSI speaking test.

One point emphasized by the integrated skills literature is the need to consider scoring all the skills that are of interest in the target language. Until recently, the trend had been to score only the spoken or written performance and not the receptive skills involved (listening or reading) (Plakans and Gebril, 2012). Without scoring all the language skills involved, reported scores on a task that appears to reflect an authentic integration of multiple skills may provide more limited information than it could. An oral interview integrates listening and speaking, with a test taker needing to comprehend the questions by using listening skills in order to answer them orally. Although interview tasks have been used to assess speaking for many decades, until recently the necessary listening skills have usually played only a small role in defining the construct and scoring the performances for these tasks.

One approach is to score the appropriateness of the response to the question, i.e., the degree that the test taker comprehends the question. This approach is used in the Center for Applied Linguistics' BEST Plus oral language test, which scores the appropriateness of the response to each question.[2] In other cases, an oral assessment may contain tasks that appear to focus more on listening than on speaking. For example, in the interview section of the FSI test, a test taker needs to understand substantial spoken language in order to then report it in English to the interviewer. In this case, perhaps a reported score could also include the skill of listening, based on the completeness or accuracy of reporting in English what was heard in the target language. An example of a rubric that captures this sort of content-responsible listening/speaking can be found in the iBT TOEFL integrated speaking task scale.[3]

[2] See http://www.cal.org/aea/bp.

[3] See https://www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf.

ADDING TARGET-LANGUAGE WRITING AS A RESPONSE MODE FOR SOME READING OR LISTENING TASKS

Writing appears to be part of the target language use domain for Foreign Service officers, although it is not included in the current FSI test. As noted in the discussion of the FSI context (see Chapter 2), writing appears to have become increasingly important in recent years, with short interpersonal written exchanges in text messages, email, or social media replacing verbal exchanges that used to take place by telephone or during in-person meetings. Further analysis (using one of the methods discussed in Chapter 3) would be needed to understand how Foreign Service officers currently need to use writing. Such a review might suggest adding a writing component to better reflect job-related language competencies.

There are a variety of ways that writing could be included in the FSI test. One example might be to develop tasks that involve writing email messages in response to reading and listening texts that are other emails or voicemail messages in the target language. Such tasks could be modeled on typical email correspondence by Foreign Service officers, perhaps involving responses to email requests, emails to request information about some topic described in a reading or listening text, or emails to briefly describe some aspect of U.S. culture or current policy in response to a written inquiry. This extension of the reading or listening tasks, with writing in the target language, could be considered an integrated skill assessment. For such an addition, FSI would need to consider how performances would be appropriately scored, as noted above in relation to assessing listening with speaking.

ADDING PAIRED AND GROUP ORAL TESTS

In recent years, researchers have explored the use of paired and group oral tests as a complement to one-on-one interview-style speaking tests. Paired and group orals were created to capture test-takers' interactional competence (Ockey and Wagner, 2018a; Roever and Kasper, 2018), allowing raters to judge whether test takers are able to comprehend the other speaker and respond appropriately, are aware of other speakers' roles in a conversation, and are able to manage conversational turn-taking, repair conversational breakdowns, and co-construct topics.

Paired oral tests resemble naturalistic conversation, mirror pair work that is common in communicatively oriented and task-based language classrooms, and can help in the measurement of interactional competence (Ducasse and Brown, 2009). Group orals generally involve three to four candidates, with the goal of eliciting group interaction. Groups are normally given 8 to 10 minutes to discuss a given topic, task, situation, or scenario, and thus group orals are more often used with test takers who are already conversant in the language (Winke, 2013).

These interactional skills are likely part of the Foreign Service target language use domain. However, there could be practical challenges to coordinating opportunities for paired or group oral tests for FSI, particularly for languages with few test takers, and there are potential fairness concerns raised by the variability of the pairings.

Paired and group orals provide challenges related to interlocutor variability that are not present in one-on-one interviews because peer testing partners are not working off scripts and may come to the test with different language proficiencies, as well as variations in personality, motivation, and background (Ockey, 2009). Research has found that individuals who are assertive, disagree, or have a self-centered approach to the speaking task can influence how other speakers are able to participate in the conversation (Lazaraton and Davis, 2008). Raters may then struggle to determine a score that is fair to a candidate who they perceived had been disadvantaged by a particular pairing (May, 2009). It is important to note, however, that pairings of candidates at different proficiency levels might not necessarily influence the resulting scores (Davis, 2009).

Paired and group oral tests have been used in a variety of high-stakes testing programs. The first high-stakes paired oral assessment was introduced in 1991 by the Cambridge English for Speakers of Other Languages Certificate of Advanced English test (Lazarton and Davis, 2008). Since then, the majority of Cambridge tests have had a paired format (Norton, 2005). The English placement test for nonnative speakers of English at Iowa State University also uses paired oral assessments as a complement to a one-on-one oral interview. The paired task involves listening to audio recordings of two speakers providing different positions on a topic, followed by an opportunity for the test takers to summarize, discuss, and defend one of the positions. The scoring criteria include specific attention to interactional competence, with consideration to connecting one's own ideas to a partner's ideas, expanding on a partner's ideas, making relevant comments, taking turns appropriately, asking appropriate questions, disagreeing politely, and answering questions in an appropriate amount of time.

The Spoken English Test of the National College English Test (the "CET-SET") in China includes a high-stakes, standardized, group-oral assessment (Zhang and Elder, 2009).[4] In the test, three to four students perform individual warm-ups with the examiner and then present monologues to the group. The students have two group discussions—one on the

[4] This test is an optional, additional component of the College English Test (CET) taken by a small number of the approximately 10 million examinees annually who have already passed the main CET (at levels 4 or 6) at universities throughout China.

presentations they gave and one on a new topic—with the test examiner then posing additional questions. The scoring criteria consider whether the candidates contribute actively to the group discussion.

INCLUDING LISTENING TASKS WITH A RANGE OF LANGUAGE VARIETIES AND UNSCRIPTED TEXTS

A review of the current FSI test using a principled approach would consider the extent to which the results generalize from the test situation to real-world circumstances. The current speaking test includes listening in the target language that is spoken by the tester in a relatively structured exchange, but daily exchanges will likely include a much wider variety of types of speech. Two of the most salient differences in spoken language to consider are language varieties and the scriptedness of text.

With respect to varieties, language can vary due to many factors, such as geographical region, education level, and ethnicity. As noted in Chapter 3, recent research has heightened appreciation of the many varieties of language that are used in natural settings. This factor can be particularly important with respect to listening comprehension, since spoken language often reflects a set of differences, including accent, that are often not present in written language. In many contexts, the dominant or national language might be a second language for many residents of that country or region, and thus accented varieties of the target language will be part of the target language use domain for Foreign Service officers in such contexts.

Research on accents shows that multiple exposures to and familiarity with a particular accent generally leads to increased comprehension of that accent (Gass and Varonis, 1984). A proficient listener in a language should be able to comprehend multiple variants of the target language and accommodate or adapt to unfamiliar accents (Canagarajah, 2006). The research is clear that a speaker can be comprehensible even with a perceived accent (Isaacs, 2008).

With respect to scriptedness, a Foreign Service officer's language use probably typically includes both scripted language, such as political speeches, and unscripted spoken language, such as informal conversations and interviews. Research on scriptedness shows that unscripted spoken texts differ from scripted spoken texts in a number of ways; listeners vary in their ability to comprehend unscripted spoken language, based in large part on their exposure to it (Wagner, 2018).

A review of the current test might identify the importance of assessing language proficiency with respect to a range of language varieties and with both scripted and unscripted varieties. By using recorded speech, the FSI test could include listening tasks with a typical set of varieties of the language Foreign Service officers may be exposed to and a mix of scripted and unscripted texts. When selecting listening texts, it is important to include whatever language varieties are prevalent in the target language use domain for that particular Foreign Service setting. For example, the listening test of the TOEFL iBT includes U.S., British, New Zealand, and Australian varieties of English. Such an expanded range of listening tasks would add additional time and expense to the testing process. However, in addition to providing better coverage of the target language use domain in some contexts, it would also likely have beneficial effects on instruction to provide test takers with exposure to the relevant range of language varieties.

INCORPORATING LANGUAGE SUPPORTS

In many situations, real-world language proficiency often involves the use of language supports (Oh, 2019). Traditional language supports include translation dictionaries, spelling and grammar checks, online dictionaries, and translation and interpretation apps, such as Google Translate. It is likely that a full review of the target language use domain for Foreign Service officers will show a number of ways that they incorporate language supports. Some situations, such as composing an email, allow the use of a language support while performing the task itself; other situations, such as having a conversation or making a presentation, may allow the use of a language support only beforehand, while preparing for the task.

There is considerable research relating to the value of providing supports to test takers, including research on the use of computers for test administration (e.g., Schaeffer et al., 1998) and extensive research on providing accommodations to students with disabilities (e.g., Luke and Schwartz, 2007). A general finding in this research is that a supporting tool may reduce the demand for some particular knowledge or skill on the test (such as foreign language vocabulary), while also adding a demand for knowing when and how to use the tool. As a result, it is important that test takers be familiar with a particular support and how it can be used.

It would be possible to incorporate the use of language supports into the FSI test with small modifications to the current tasks. For example, in the work-related exchange task in the current speaking test, the test taker could be allowed to use a translation dictionary during the initial preparation time to look up key vocabulary, as one would likely do in the target language use domain. In the in-depth task on the reading test, the test taker could be allowed to use a translation dictionary or an app to look up vocabulary or phrases. In both of these examples, the test administration and scoring of the current test would be substantially unchanged, but the interpretation of the result would be subtly changed to include the ability to effectively use language supports. It would also be possible to develop new tasks that allow test takers to use translation apps to accomplish a particular task in ways that resemble the way they are typically used in the real world.

ADDING A SCENARIO-BASED ASSESSMENT

As noted above, the FSI test already uses tasks that draw on several language skills and that resemble aspects of common tasks that many Foreign Service officers need to perform—for example, the need to build understandings through interaction in the target language and report those understandings in English. Assessments that use richer "scenarios" could further broaden the way that the assessment tasks reflect real-world competencies, such as writing informational reports. Scenario-based approaches to assessment are motivated by domain analyses, which show that real-life tasks are very different from those used in traditional language assessments (e.g., Griffin et al., 2012; National Research Council, 2001b; Organisation for Economic Co-operation and Development, 2003, 2016, 2018). Many language assessments give test takers a series of unrelated tasks, each focusing primarily on a single language skill, such as reading or listening, which do not reflect the complexity of language use in real-world goal-oriented activities. Another example is when Foreign Service officers need to work collaboratively in teams to gather information, discuss an issue, propose a consensus-based solution to a problem, and share their solution with other colleagues.

Such real-life scenarios can be used as the basis for richer assessment activities that reflect the language proficiency needed to collaboratively solve problems in another language. As described above (Chapter 3), the work-related task in the current FSI speaking test allows the test taker to spend 5 minutes to prepare an initial statement on a selected topic, which is then followed by an exchange with the tester. This task could be enriched as a scenario in numerous ways. For example, rather than having the test taker invent a hypothetical statement, the task could provide several short readings to use as the basis for the statement, requiring the test taker to build an understanding of the issue through these documents, and then present and discuss the issues in the target language. The task could be further extended by having the test taker write an email in the target language that summarizes the key points raised during the discussion. Depending on the specific scenarios that might be relevant to the Foreign Service context, the initial readings could be in the target language, in English, or a mix. Such an enriched task could provide a demonstration of language proficiency that relates more closely to the kind of tasks often carried out by Foreign Service officers. Adding a scenario-based assessment activity would require significant change to the test administration.

There are examples of the use of scenario-based assessment in both education and employment settings. On the education side, significant research has been carried out at the Educational Testing Service to examine new ways of improving assessments in K–12 mathematics, science, and English language arts in the United States.[5] Perhaps the most well-developed, scenario-based assessments in education are the international PISA tests,[6] which are designed to determine if 15-year-old students can apply understandings from school to real-life situations. PISA test takers are not asked to recall factual information, but to use what they have learned to interpret texts, explain phenomena, and solve problems using reasoning skills similar to those in real life. In some PISA tests the problem solving is collaborative.

An example of scenario-based assessment specifically related to language proficiency is the placement test in English as a second language that is being developed for the Community Language Program at Teachers College in New York City (Purpura, 2019; Purpura and Turner, 2018). The overarching goal of the intermediate module is for test takers to make an oral pitch for an educational trip abroad to a selection committee on behalf of a virtual team, which requires a coherent series of interrelated subtasks involving multiple language skills that need to be performed on the path toward scenario completion ("pitch the trip").

On the employment side, scenario-based approaches to assessment are often used for tests related to hiring, placement, and promotion. These approaches range from situational judgment tests to more realistic work samples and exercises that place candidates in situations that reproduce key aspects of a work setting to gauge their competence related to interpersonal skills, communication skills, problem solving, or adaptability (Pulakos and Kantrowitz, 2016). Even testing formats that do not strive for such realism, such as structured interviews, can be designed to include the use of scenarios that ask candidates what they would do or did do in a particular situation (Levashina et al., 2014). One particularly relevant example of a scenario-based, high-stakes employment test is the State Department's own Foreign Service Officer Test, which includes a situational judgment component.

[5] See the CBAL® Initiative (Cognitively Based Assessments of, for, and as Learning) at https://www.ets.org/cbal. Also see the Reading for Understanding Initiative at https://www.ets.org/research/topics/reading_for_understanding/publications.

[6] PISA, the Program for International Student Assessment is a worldwide study by the Organisation for Economic Co-operation and Development. It was first administered in 2000 and has been repeated every 3 years.

INCORPORATING PORTFOLIOS OF WORK SAMPLES

The committee's statement of task (see Chapter 1) specifically asks about the possibility of using portfolios in FSI's test. Portfolios are often discussed in the educational literature in the context of collections of student work that are assembled to provide evidence of competency to complement or replace a formal assessment.[7] Portfolios are also sometimes discussed in the context of collections of employee work (Dorsey, 2005). For FSI, either of these uses might be considered, with evidence of language-related work assembled during language instruction or on the job.

Portfolios have the potential to provide information about a broader range of language performances than can be sampled during a short formal assessment. In the case of job-related work samples requiring use of a foreign language, such information would clearly relate to the target language use domain because the work samples would be drawn from the domain. For FSI, a portfolio could be used in addition to the current test, which would be an example of using multiple measures for decision making. Portfolios may help address concerns that some test takers may be able to pass the test but not actually be able to use the target language in their work, while others may be able to use the target language in their work but not pass the test.

The weaknesses of portfolios relate to the difficulty of interpreting the information they provide about what a test taker knows and can do: they can be hard to standardize and can be affected by factors that are difficult to control, such as the circumstances under which the included performances were obtained and the extent of assistance provided to the test taker (Brown and Hudson, 1998). The use of portfolios as the basis for making high-stakes decisions about test takers has raised questions about the legitimacy of a selected portfolio as an accurate reflection of a test-taker's ability to respond independently, the reliability and generalizability of the scores, the comparability of portfolio tasks across administrations, and unintended effects on instruction when instructional activities are designated for inclusion in a portfolio (East, 2015, 2016; Herman et al., 1993; Koretz, 1998; National Research Council, 2008). Portfolios can also be time consuming to prepare and score.

Nonetheless, portfolios have been used in a variety of educational settings related to language instruction. An example is the Council of Europe's *European Language Portfolio*, which was intended for individuals

[7] In the context of education, student portfolios often also include other information in addition to work samples, such as students' self-assessment, learning history, or learning goals (Cummins and Davesne, 2009). These are likely to be important for future educational decisions, but they are not considered here because they are not relevant to making high-stakes decisions about an individual's level of competency.

to keep a record of their language learning achievements and experiences, both formal and informal (Little, 2011). In Canada, portfolios are used as the final test in the government's language instruction courses for immigrants and refugees, which are a required step in the immigration process (Pettis, 2014). In New Zealand, portfolios can be used in place of an oral proficiency interview at the end of term for high school students in foreign language courses (East and Scott, 2011). One example of the use of portfolios in a work context is the National Board for Professional Teaching Standards in the United States, which uses structured portfolios related to a video of a lesson as part of the process for advanced certification for K–12 classroom teachers (National Research Council, 2008).

ADDING A COMPUTER-ADMINISTERED TEST USING SHORT ASSESSMENT TASKS

Computer-administered tests with large numbers of short assessment tasks are widely used (Luecht and Sireci, 2011). For example, a number of tests of English language for nonnative speakers assess reading and listening using multiple-choice comprehension items, such as the TOEFL iBT,[8] the International English Language Testing System,[9] the PTE Academic,[10] and the WIDA ACCESS for K–12 students.[11] The Defense Language Proficiency Tests[12] use a similar approach to assess reading and listening in foreign languages based on the ILR framework. In addition to multiple-choice questions, some of these tests use other selected- or constructed-response formats, such as matching, diagram labeling, fill-in-the-blanks, sentence completion, short answer, highlighting the correct summary, selecting a missing word, and highlighting incorrect words.

Such a test might be considered in response to a goal of broadening the current test coverage of Foreign Service topics. For example, the current FSI test is intended to assess a test-taker's ability to understand and use professional-level vocabulary, discourse, and concepts in relation to a range of political, economic, and social topics that are relevant to the Foreign Service; however, only two or three reading texts are used for in-depth reading. A test using short assessment tasks in reading or listening could sample from a greater range of discourse and topics than can the limited number of reading passages sampled in the current test. Expanding the

[8] See https://www.ets.org/toefl/ibt/about.

[9] See https://takeielts.britishcouncil.org/take-ielts/what-ielts.

[10] See https://www.examenglish.com/PTE/PTE_Academic.htm.

[11] This test is most often used as a screening test to determine the language level of students entering a school system.

[12] The Defense Language Proficiency Tests are foreign language tests produced by the Defense Language Institute–Foreign Language Center and used by the U.S. Department of Defense.

breadth of coverage also has the potential to yield more information about the extent to which a test taker can understand a wide breadth of professional vocabulary, discourse, and concepts in the FSI target language, thus improving the reliability and generalizability of scores. However, unlike the current reading test, a test using many short assessment tasks might provide linguistic breadth by sacrificing communicative depth that reflects real-life language use in the domain. A computer-administered test using selected response questions would limit the ability to probe the test-taker's responses, in contrast to the current reading tests. In addition, initial development costs for computer-administered tests would need to be considered; high development costs could affect their practicality for low-frequency languages.

In recent years, there has been growing interest in computer-adaptive tests, which reduce the time spent on questions that are clearly too easy or too difficult for the test taker, focusing instead on questions that appear to be at the border of the test-taker's ability (e.g., Luecht and Nungester, 1998; Van der Linden and Glas, 2000; Wainer et al., 2000). The FSI test already includes adaptation to the level of difficulty appropriate to the test taker. In the speaking test, this adaptation occurs as a result of training in the certification process, as the tester modulates the level of speech that is used. In the reading test, the adaptation occurs explicitly in the choice of longer and more linguistically complex reading passages at a particular level of difficulty after the test-taker's performance on the shorter reading passages. A computer-adaptive test could potentially implement such adaptation in a more standardized way.

The responses on computer-adaptive tests are automatically scored in real time, which allows the scores on prior questions to guide the choice of the questions that follow from a pool of possible questions. Although reading and listening potentially lend themselves to computer-adaptive testing because of the frequent use of machine-scorable questions, the approach has not been widely embraced in language proficiency testing because of the cost involved in developing and calibrating the necessary pool of items. Because of this requirement, the approach is feasible only for large-scale tests. However, this extra expense can be limited by using a "multistage" adaptive approach in which short sets of questions are administered and scored in a series of stages. Performance on one set of questions is used to select the next to administer to a given examinee. Generally, this approach reduces the size of the item pool required (e.g., Leucht et al., 2006; Yamamoto et al., 2018; Yan et al., 2014; Zenisky et al., 2010). For FSI, the small numbers of test takers for many languages may still make the development of computer-adaptive approaches impractical and prohibitively expensive.

A simpler but conceptually related approach would make use of a "two-step" process. In this approach, a screener test would be used to estimate if test takers are likely at or above a threshold level of proficiency that would enable them to achieve the desired proficiency rating of 3/3 (or higher) on the full FSI test. Test takers below this threshold would not go on to take the full test, and the score on the screener would be their official score. For expedience and cost-effectiveness, the screener test could be computer administered and consist of questions with machine-scorable response formats. Moreover, the screener could contain a machine-scorable listening component that may predict oral language performance (i.e., speaking) on the full test.

USING AUTOMATED ASSESSMENT OF SPEAKING

Recognizing the intense resources that are currently being devoted to developing artificial intelligence (AI) techniques related to language, the committee highlights a possible change that would be explicitly forward looking: the use of automated scoring for the assessment of speaking. Unlike the other changes discussed, the possibility of a change to (or adoption of some elements of) automated scoring depends on larger breakthroughs that are being pursued by computer science researchers and developers. The intent of including this possibility on the list is to highlight the potential value of new technologies that may become available in a decade or so in order to sensitize FSI to these future possibilities.

Technology-based speaking tests are currently used routinely in some large testing programs to elicit test-takers' speech in response to recorded prompts. The test-taker's recorded responses are typically rated later by two raters. The TOEFL iBT is an example. It includes four speaking prompts that are recorded and later scored by human raters. The computerized version of the ACTFL Oral Proficiency Interview (see Isbell and Winke, 2019) and the now decommissioned Computerized Oral Proficiency Test from the Center for Applied Linguistics (see Malabonga et al., 2005) are two other examples of language tests that have been used for a range of world languages that collect responses to recorded prompts that are later scored by human raters. Although such computer-based tests can often provide more standardized assessment tasks than face-to-face interviews, they may show more limited aspects of language than face-to-face interactions (Kenyon and Malabonga, 2001; Quaid, 2018). In addition, other features of oral communication are not addressed by computer-based test tasks, such as the construction of meaning, social sensitivity, the conveyance of empathy, and turn-taking. While face-to-face interviews and computer-mediated platforms might yield comparable scores statistically with respect to basic features of language use, it is likely that the different modes of testing are tapping different skills (Qian, 2009).

One goal of automated scoring is to use scoring engines for the recorded speech from technology-based assessments (Wang et al., 2018).

The automated score may take the place of one of two human raters, for example, reducing costs, and interrater reliability could be calculated as between the automated score and the human score, with a second human rater only needed when the two do not agree.

Some operational tests already use limited AI to produce automated scores and do not involve human raters. Pearson's Versant (formerly owned by Ordinate and called the PhonePass Test) takes 15 minutes and is automatically scored. The automated scores are from elicited imitation, sentence repetition, and short spoken responses, which are speaking tasks scored through careful elicitation and do not involve authentic communication (Chun, 2008). Pearson's PTE Academic is automatically scored as well: test takers read text aloud, repeat sentences, describe images, and provide brief responses to questions. We note, however, that these types of automatically scored tests have been criticized as inauthentic, underrepresenting the speaking construct, and not assessing real conversation (Chun, 2008).

Despite the limitations of technology-based speaking tests and automatically scored speaking tests, there is a growing body of research on human conversation with chatbots and virtual assistants that is helping to inform and advance a set of AI technologies related to conversation (Ciechanowski et al., 2019). Computer-human interaction is the ultimate goal of this branch of AI assessment of speaking. This goal will be achieved when the computerized AI voice can ask questions and guide the direction of the conversation based on test-takers' responses in real time. For example, several testing companies are researching or using AI to rate computer-based tests' recorded speech samples and limited conversations (Chen et al., 2018; Ramanarayanan et al., 2020). Such ratings could form one part of a technology-based conversational system, although AI techniques cannot yet reliably score important qualities of human interaction, such as pragmatic appropriateness, collegiality, and humor (Bernstein, 2013). Future AI breakthroughs could substantially improve the capabilities of such systems with the potential of making technology-based oral testing more interactive.

As these technologies continue to be developed, they offer the possibility of greater standardization and reduced cost in the administration and scoring of speaking, while preserving more of the elements of human conversation that are missing from current technology-based speaking tests. Thus, at some future time, such systems could be attractive for use in FSI's test.

PROVIDING TRANSPARENT SCORING CRITERIA

Because of the subjective nature of scoring extended responses, such as those elicited by the FSI test, it is important that scorers be well trained to apply the criteria laid out in the scoring rubric and that the criteria clearly reflect the knowledge, skills, and abilities that are assessed. Rubrics make the scoring less subjective, helping scorers to reliably and fairly transform a test-taker's performance into a score by using agreed-upon criteria. The body of research on developing effective scoring rubrics for writing and speaking is sizable (for an overview, see Van Moere, 2013).

In addition to developing scoring rubrics, testing programs need to provide scorers with extensive and ongoing training to use the rubrics consistently. Initial training related to the meaning of the different criteria included in the rubric is important, but scorers also need regular norming and recalibration to correct for drift and to ensure that their scores are consistent with those given by other scorers. There is a considerable amount of guidance for rater training procedures in language assessment (e.g., Van Moere, 2013; Weigle, 1998). Scoring rubrics and rater training procedures need to give particular attention to the scoring of different varieties of the language, which can be particularly challenging when test takers and raters may come from a range of language backgrounds.

Scoring rubrics are generally publicly available as part of the documentation provided by a high-stakes testing program (see discussion of professional testing standards in Chapter 6). To help ensure the reliability, fairness, and transparency of the scoring process used in the FSI test—as well as the perception of that reliability and fairness—FSI should consider providing more information to the test takers and users about its scoring rubrics and procedures, as well as its scorer training processes. Transparent scoring rubrics can also improve performance by better aligning teaching and learning with valued outcomes (Jonsson, 2014; Tillema et al., 2011). Providing more transparent scoring criteria could be part of an overall effort to develop a shared understanding about language assessment across all stakeholders in the State Department.

USING ADDITIONAL SCORERS

One source of variability in the FSI test relates to the tester and the examiner who administer the test. These two individuals serve both as interlocutors—to prompt and collect the language performance on the test—and as scorers of that language performance. Without adding any complexity to the administration of the test, FSI could use the video recording of the test for a separate scoring by a third independent scorer. Such a review by a third scorer is currently used by FSI whenever scores are challenged by a test taker. However, if there are concerns about the reliability or fairness of the current test procedure, a rating by a third scorer could be added as a regular feature of the FSI test. This addition would reduce the effects of any one scorer on the outcome of the test, and it would have the additional benefit of providing regular information about the consistency in the ratings across scorers. The value of additional scorers, whether routinely or for systematic samples, can be examined quantitatively with a research study before an implementation decision is made.

Another version of this possible change could involve changes to the scoring procedure so that the FSI tester and examiner provide their scores independently. The current scoring procedure starts with the tester and examiner reaching consensus about an overall holistic score before separately developing index scores that reflect their independent evaluations of the five factors. This scoring procedure could be altered so that the tester and examiner provide scores separately before reaching consensus. An examination of the ratings awarded independently would provide information about the range of difference between the two scorers, which could be monitored to provide additional information about the reliability of the scoring process.

PROVIDING MORE DETAILED SCORE REPORTS

One aspect of a testing program that needs to be considered in evaluating its validity is the different ways the test results are interpreted ("meaning of scores") and then used, and the resulting consequences of those uses on everyone involved in the program. Substantial recent research demonstrates the value of providing more meaningful score reports (e.g., Hambleton and Zenisky, 2013; Zapata-Rivera, 2018).

For FSI, if there is limited understanding on the part of test takers and score users about the criteria used for scoring test-takers' performances, additional information could be provided. For example, providing more information than a single ILR level in the score report might be useful because it allows a more comprehensive understanding of what the scores mean. Additional information in the score report could help test takers understand the criteria that are being applied across all test takers during scoring. If the review of FSI's testing program shows any potential concerns about its fairness, additional transparency about the reasons for individual scores can help address those concerns, as well as help identify aspects of the scoring process that may need to be improved. As with the comments above about transparent scoring, the provision of more detailed score reports could be part of an overall effort to develop a shared understanding about language assessment across all stakeholders across the State Department.

5

Interpreting FSI Test Scores

Building on the discussion in Chapter 4 of possible changes to the current FSI test and its scoring that might be motivated by a principled approach review of the test, this chapter considers the way that the FSI test scores are interpreted. A key element of this consideration is the role played by the skill-level descriptions of the Interagency Language Roundtable (ILR) framework.

THE ROLE OF THE ILR FRAMEWORK

FSI and many other government language testing programs use the skill-level descriptions of the ILR framework to understand language proficiency across all levels for all languages. Because the descriptions are used in so many different ways as a foundation for government language testing programs, it can sometimes be difficult in the government context to see the distinction between different aspects of assessment programs as shown in Figure 1-1 (in Chapter 1).

In government testing programs, the ILR framework is used as a substitute for a detailed description of the target language use domain of interest for a specific test. However, as discussed in Chapter 2, a full understanding of the target language use domain for any specific government language use requires more domain-specific detail than is included in the ILR skill-level descriptions.

For example, Foreign Service officers need to use the target language to engage in social conversation and write informal email messages, in addition to understanding formal presentations, which are different language uses from those that an analyst in one of the intelligence services might need. It is important for the FSI testing program to develop a detailed understanding of language use that is specific for Foreign Service officers. Since the ILR framework describes language proficiency broadly, its descriptions are not sufficiently detailed to design a test for a specific purpose and build a solid validity argument for that test (see Chapter 2).

The use of the ILR framework can also obscure the distinctions between the ILR level scores awarded on the FSI test, their interpretation, and their use to make decisions. Because the FSI test is scored in terms of the ILR skill-level descriptions, which have defined interpretations and known uses in making decisions for Foreign Service officers, it can appear that there is no distinction between the score awarded on the test and its interpretation and subsequent use. Yet scoring, interpretation, and use are distinct, as shown graphically in Figure 1-1 (in Chapter 1):

- The *score* on the test reflects an evaluation of a specific test-taker's performance on specific test tasks based on a set of skill-level descriptions.
- The *interpretation* of the score involves a generalization from the language proficiency elicited in the test and evaluated through the descriptions to the test-taker's proficiency in the real world.
- The *uses* that flow from score interpretation involve decisions that reflect the adequacy of the test-taker's inferred language proficiency to function meaningfully and appropriately in the target language use domain.

Fundamentally, the ILR framework provides a way for many government testing programs to interpret a test score in terms of what the government considers general or functional language proficiency and to link that interpretation to a set of personnel decisions, with related consequences. The ILR framework makes it possible to discuss personnel policies related to assessment of employees' language proficiency in common terms across government agencies. As described in Chapter 3, most language-designated positions for Foreign Service officers are specified as requiring certification at the ILR level 3 in both speaking and reading. That certification is a requirement for long-term retention in the Foreign Service and is linked to incentive pay. The corresponding personnel policies of other government agencies with assessment of employees' language proficiency are similarly described with respect to the levels defined within the ILR framework.

However, as a widely used framework across the government, the ILR framework cannot fully specify the necessary assessment details that are specific to FSI's context and purpose. The importance of these details is highlighted in Figure 1-1 by the ring related to understanding of sociocultural and institutional contexts. For example, the ILR framework does not incorporate the details about professional-level Foreign Service vocabulary that are reflected in the assessment tasks and topics used in the FSI test and the underlying scoring process used to evaluate performances on those specific tasks. Similarly, although the ILR framework provides examples of different levels of language proficiency, it does not reflect critical language uses for which a test-taker's language proficiency are being inferred. Finally, although the ILR framework is used in Foreign Service personnel policies that affect retention and pay decisions that can be compared across government agencies, it does not specify the kinds of mission-critical consequences that could occur with Foreign Service officers in the field who do not have adequate language proficiency for their positions (see Box 3-1 in Chapter 3).

MAPPING THE FSI TEST TO THE ILR FRAMEWORK

As explained in Chapter 1, a principled approach to test development will rest on a detailed understanding of the target language use domain, tasks that elicit performances that reflect key aspects of the domain, clear rules for scoring those test performances, and interpretations of those scores that lead to inferences about language proficiency in the domain. In the current FSI testing program, each of these aspects is described in terms of the ILR framework—rather than the target language use domain for Foreign Service officers. As a result, the entire testing program is geared toward producing a result that can be compared with other government testing programs based on the ILR framework.

A shift in focus to the target language use domain has the potential to strengthen the FSI test. However, this shift would mean that many aspects of the assessment would rest on the target language use domain in the Foreign Service, which may not be specifically addressed in the ILR framework. With such an approach, the ILR framework could retain its essential role in helping coordinate personnel policies across government agencies that assess employees' language proficiency, but FSI's testing program would not necessarily be defined solely in terms of the ILR framework. The testing program could use a more detailed and specific understanding of the target language use domain in the Foreign Service as the basis for designing tasks, scoring test-taker performances on those tasks, and interpreting those performances with respect to the required language proficiency of Foreign Service officers.

In some ways, the FSI testing program already elaborates its understanding of the ILR framework to consider the target language use domain for the Foreign Service, especially with respect to the specific tasks in the speaking test, which is different from the more common oral proficiency interview used in other government agencies. However, explicitly acknowledging that the understanding of the target language use domain is driving the test raises the possibility of providing numerical scores for the test that are not directly described in terms of the ILR skill-level descriptions. In this approach, it may be necessary to map the resulting test scores to the ILR framework to link to the common personnel policies across government agencies.

For example, suppose FSI decided to augment its current speaking test with a technology-based test using many short listening tasks (see Chapter 4) that are scored correct or incorrect. This new test might result in a score continuum of 0 to 60 points, bearing no relation to the levels of the ILR framework. The results of this new test would need to be combined in some way with the results of the current speaking test to produce an aggregate result. One way to do this might be to simply add the score from the new technology-based test with the 120-point "index scale" that is produced (though not specifically used or reported) during the scoring process of the current speaking test. Or the combination could reflect different weights or thresholds, depending on the meaning of the two different scales. For either approach, or any other, the resulting aggregate numerical score would still need to be mapped to the levels of the ILR framework.

There are well-developed procedures for carrying out such mappings and providing evidence in support of interpreting performances on an assessment in terms of an external set of proficiency levels (such as the ILR skill-level descriptions). One way of performing the mapping is by a "contrasting groups" (or "examinee-centered" or "empirical") approach, in which test takers with known ILR level scores from the current test would be given the new test as well (see, e.g., Livingston and Zieky, 1982; see also, e.g., Cizek, 2012; Cizek and Bunch, 2007; Hambleton and Pitoniak, 2006). By having a set of test takers take both tests, it would be possible to understand the relationship between the numerical scores on the new test and the ILR level scores from the current test. This information could then be used to map the numerical scores on the new test to the ILR skill-level descriptions for policy purposes—such as personnel decisions—in a way that would produce similar numbers of examinees reaching those levels as the current test.

Another way of mapping from the scores of the new test to the ILR skill-level descriptions would be by using standard-setting (or "test-centered") processes, which use groups of qualified panelists to go through a standardized procedure with a test's tasks to define one or more cut scores between different levels of performance (see, e.g., Cizek, 2012; Cizek and Bunch, 2007; Hambleton and Pitoniak, 2006). The Council of Europe's manual for relating performances on language tests to the CEFR[1] provides instructions on how to implement this type of mapping (Figueras and Noijons, 2009).

The widely used Test of English as a Foreign Language (TOEFL) provides an example of using a "test-centered" mapping method. TOEFL has its own score scale, which test users have long used to make decisions about test takers. A major use of TOEFL iBT is to determine English-language readiness for pursuing a course of study at an English-language university. Schools, for example, may require a TOEFL iBT score of 85, of the 120 total, for international students to meet a university admission criterion related to language proficiency, though each college or university can set its own cut score for admission. However, with the increasing use of the CEFR as a framework for describing language proficiency, a number of English-speaking universities outside North America wanted to define their language proficiency admission criteria in terms of the six CEFR levels. In response, the Educational Testing Service (ETS) conducted a study to map performances of the TOEFL iBT onto the defined proficiency levels described by the CEFR (Tannenbaum and Wylie, 2008). The study was done "[f]or test users and decision makers who wished to interpret TOEFL iBT test scores in terms of the CEFR levels in order to inform their decisions" (Papageorgiou et al., 2015, p. 2). Based on the study, ETS established boundary scores in terms of scale scores on the TOEFL iBT that could be interpreted as entry scores into CEFR proficiency levels.

Although "test-centered" standard-setting processes would provide evidence for the correspondence between the meaning of the scores on the new test and the ILR skill-level descriptions, the procedures will not ensure that the new test produces roughly similar numbers of examinees achieving those ILR levels. If it is important for FSI that a new test maintain roughly similar distributions of outcomes in terms of ILR level scores, then the mapping should be carried out using an "examinee-centered" approach.

CONSIDERATIONS FOR FSI

Fundamentally, the ILR framework provides a way for multiple government testing programs to interpret a test score in terms of a common government understanding of language proficiency and to link that interpretation to a set of personnel decisions. As discussed above, however, although the ILR framework defines some of the context for the FSI testing program and the interpretation of the scores it produces, it cannot provide the full level of detail needed to design and validate a test for FSI. A prin-

[1] CEFR, the Common European Framework of Reference for Languages: Learning, Teaching, Assessment, is a guideline used to describe achievements of learners of foreign languages, principally in Europe.

cipled approach to test development for the FSI testing program will rest on a detailed understanding of language and the target language use domain and the sociocultural and institutional contexts of the Foreign Service; assessment tasks that elicit performances that reflect key aspects of the target language use domain; scoring that fairly and reliably evaluates those test performances; and interpretations of those scores that lead to appropriate inferences about language proficiency in the target language use domain for the purpose of making decisions about Foreign Service officers.

As FSI uses principled approaches to understand its current test and consider possible changes to it, it may be worth considering approaches to scoring that are based on a scale score that are not so directly linked to the skill-level descriptions of the ILR framework. It would still be possible to maintain the framework's role in coordinating language proficiency personnel policies across government agencies by mapping a new FSI test score to the skill-level descriptions. There are a variety of techniques to setting cut scores that can be used to perform such a mapping.

6

Evaluating Validity in the FSI Context

This chapter addresses a framework for an ongoing evaluation of the FSI test, which fundamentally relates to the validity of its scores. It follows the above chapters that discussed the elements that relate to the development of an assessment program: the understanding of language, the sociocultural and institutional contexts, and the target language use domain (Chapters 2 and 3); the tasks, the performances they elicit, and the resulting scoring of those performances (Chapter 4); and the interpretation of those scores that supports their use (Chapter 5).

An ongoing evaluation of the FSI test will need to consider such questions as the following:

- Do the results of the assessment mean what the test designers think they mean for the context in which the assessment is used, and does the assessment function in the way they intended?
- Are the interpretations of those scores useful, appropriate, and sufficient to make the high-stakes decisions that are made by FSI?
- Are the consequences of those decisions beneficial to everyone involved in the testing program and, overall, to the Foreign Service?

In Figure 6-1 (which reproduces Figure 1-1, in Chapter 1), these questions are captured by the arrows, which highlight the relationships among the contexts and elements of the assessment program.

It is important for any testing program to articulate the claims that a test is intended to support and to document the validity of these claims with empirical evidence. This evidence should also include information

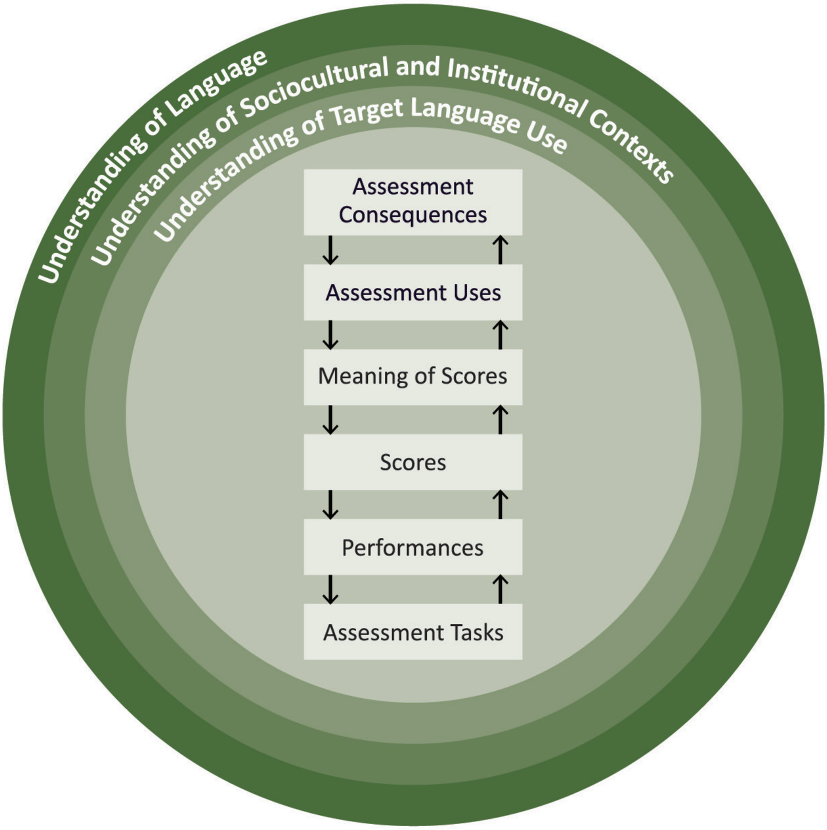
Understanding of Language
Understanding of Sociocultural and Institutional Contexts
Understanding of Target Language Use
Assessment Consequences
Assessment Uses
Meaning of Scores
Scores
Performances
Assessment Tasks

FIGURE 6-1 A principled approach to language assessment design and validation.

about the technical qualities of the test scores, such as the extent to which they are reliable and generalize across assessment tasks, assessment occasions, and scorers. This evidence should also include results from studies of fairness, including analyses to check whether test performance is biased by any factors that are irrelevant to the test's construct, such as the test-taker's gender, age, race, or ethnicity. Evidence should also be collected about the consequences of the use of the test on the decisions the test is used to support, on the test takers, and on others in the Foreign Service.

The evidence about the test, its use, and its consequences should be considered in light of the current understanding of language, the sociocultural and institutional contexts of the testing program, and the target language use domain. This consideration of evidence related to the test and its functioning is not a one-time effort but needs to be an ongoing program of research. Over time, there will be changes in the understanding of language, the contexts of the testing program, and the target language use domain, so there needs to be continuing effort to consider whether a test is working as intended.

This chapter first discusses examples of claims about performance on the FSI tests and strategies for evaluating their validity. It then provides an overview of relevant professional standards, which provide some generally applicable best practices related to the design and evaluation of assessment programs.

EVALUATING THE VALIDITY OF CLAIMS

This section offers four examples of claims made about test-takers' performance on an assessment and the strategies for validating them; it draws extensively from Bachman and Palmer's (2010) *Language Assessment in Practice*. It is not intended to be an exhaustive discussion of the process of evaluating the validity of test-based inferences but rather to provide a few concrete examples of the kinds of issues that need to be investigated during test validation.

Example Claim 1: Tasks and Performances

The first example claim relates to the tasks that are used in the assessment and the performances they elicit, as well as the target language use domain (the bottom two boxes and the inner ring in Figure 6-1):

- Claim: The characteristics of the assessment tasks and the performances they elicit correspond closely to the characteristics of the tasks and the performances they require in the target language use domain (Bachman and Palmer, 2010, p. 234).

This claim is fundamental because it focuses on the relationship between the test-taker's performance in the assessment situation, which is a function of the characteristics of the assessment tasks used, and the test-taker's performance in the real world in conditions with similar characteristics. In the context of FSI, this claim concerns the relationships between the proficiency needed to address the kinds of language-based tasks that are on the FSI test and the proficiency needed to address the kinds of tasks that Foreign Service officers need to carry out in the target language.

Performance on tasks in the assessment situation is never identical to performance in the real world. To demonstrate that this claim is reasonable and valid, it is necessary to gather data that allows FSI to estimate the extent of similarities and differences between task characteristics and the performances they elicit to confirm that performance in the testing situation is likely to require similar levels of language proficiency as performance in the real world. It is also important to consider whether the key aspects of the target language use domain are represented in the assessment tasks.

One way to study the correspondence between the target language use domain and the task situations is to list the characteristics of the target language use domain tasks and the characteristics of the assessment tasks (Bachman and Palmer, 2010, p. 235). Conceptually, this correspondence exercise is simple, but it can be time-consuming to carry out. The test tasks can be readily examined by the test designer, but the understanding of the tasks and performances in the target language use domain could require a substantial data collection effort, using one of the analysis methods described in Chapter 3. For FSI, it is necessary to understand the full range of the tasks that Foreign Service officers need to carry out in the target language and their relative frequency and importance. For example, a language proficiency test that focuses on making formal presentations and reading news articles will not capture the full range of linguistic resources needed for a particular job task that primarily requires social conversation and the exchange of email messages.

Example Claim 2: Evaluating Task Performances to Produce Test Scores

The second example claim looks at the way the performances on the test are evaluated (the second and third lower boxes and the inner ring in Figure 6-1):

- Claim: The criteria and procedures for evaluating responses to the assessment tasks correspond closely to the way language users assess performance of language tasks in the target language use domain (Bachman and Palmer, 2010, p. 236).

This claim focuses on the way that performance on the assessment is evaluated in order to produce scores. Like the first claim, this claim involves comparing the test situation and the target language use domain, but here the focus is on the criteria used to assess performances. Ideally, performance on assessment tasks is evaluated with criteria that are relevant in the real world.

In evaluating this claim, it is important to investigate the similarity in the evaluation of performance between the target language use domain and the assessment. For example, consider the extent to which the use of standard grammar is valued in the target language use domain. If standard grammar is important in the real world, then it should be important on the assessment—and vice versa. However, if accuracy is not as important in the target language use domain as, say, fluency, then the scoring criteria on the test should reflect that. This sort of evaluation will also rest on an analysis of the target language use domain using one of the methods described in Chapter 3, as well as on the understanding of language used and the understanding of the sociocultural and institutional contexts of the assessment.

For FSI, it will be important to consider what kind of task performance is adequate in the target language use domain for Foreign Service officers and what features of that performance are most important. In some situations, standard grammar, vocabulary, and accent may be critical; in other situations, an ability to communicate and understand meaning with less-than-perfect language may be sufficient. The scoring criteria should reflect the priorities in the target language use domain. Also, different tasks may require test takers to engage with different audiences for different purposes, and their performance might be scored differently for each task in accordance with how they would be evaluated in the target language use domain.

This claim is one for which issues of fairness may arise, with questions related to the criteria that are used in the scoring process.[1] For example, even with an explicit scoring rubric that focuses on the communication of meaning, it could turn out that scorers primarily focus on errors in grammar and vocabulary or are strongly affected by particular accents. Studies of the scoring process using duplicate scoring by highly trained scorers might be a source of evidence about the way task performances are evaluated in the test situation.

Example Claim 3: Scores and Their Interpretation

The third example claim looks at the interpretations of the scores that are produced by the test (the middle two boxes in Figure 6-1):

- Claim: The test scores that summarize the test-taker's performance are interpreted in a way that generalizes to performance in the target language use domain (Bachman and Palmer, 2010, pp. 237–238).

This claim focuses on the way that test users interpret the scores and the inferences about test-takers' language proficiency that they believe are justified by the scores. The interpretation of a set of test scores again concerns a relationship with the target language use domain, but here the relationship is focused on inferences about the adequacy of a test-taker's language proficiency that are made based on the test score.

[1] For a comprehensive discussion of issues related to fairness, see Kunnan (2018).

For FSI, the current test is interpreted through the Interagency Language Roundtable (ILR) framework (see Chapter 5), and the interpretation suggests that test takers who receive a score of 3 or higher have adequate language proficiency to perform the tasks they will need to perform at their posts. Investigating this claim in the FSI context might involve collecting information from Foreign Service officers in the field about their ability to successfully carry out different typical tasks in the target language and comparing that information to their test scores. In other contexts, where a cut score between performance levels has been defined using a standard-setting process, investigating this claim might involve collecting evidence of the robustness of the judgments used in the standard-setting process.

Example Claim 4: Test Uses and Consequences

This claim concerns the way the test results are used and the consequences of those uses (the two uppermost boxes in Figure 6-1 and the ring related to sociocultural and institutional contexts):

- Claim: The consequences of using an assessment and of the decisions that are made on the basis of that assessment are beneficial to the test takers and to the Foreign Service generally (Bachman and Palmer, 2010, p. 178).

Tests are often used for high-stakes decisions that can have a major impact on a test-taker's life and career. Identifying the consequences of these decisions is an important part of establishing the overall validity of a specific use of a test (e.g., Bachman, 2005; Messick, 1996). For a Foreign Service officer, passing or failing a language test can affect a test-taker's salary and long-term retention in the Foreign Service, as well as a range of subjective attributes such as self-image. Beyond the test takers themselves, the use of the test to make decisions about the placement of Foreign Service officers also affects the ability of embassies and consulates around the world to carry out their work. Inaccurate decisions could result in placing unqualified personnel in overseas positions or in preventing the placement of qualified personnel in positions where they could be effective.

In addition, the FSI test affects instruction in FSI classrooms, as teachers react to the content and format of the test by adapting their teaching approaches and instructional materials to prepare their students to be successful. This instructional response is known in the field as washback. The FSI test outcomes may also be perceived by instructors and administrators as measures of the language program's instructional effectiveness if the test takers had recently completed the FSI training program, a consequence of the direct linkage between instruction and assessment. The goal should be to create a test that assesses the aspects of communicative competence that Foreign Service officers need, both to more accurately identify the language proficiency of the test takers and to encourage the language training program to develop the aspects of language proficiency that are needed.

This example claim involves a number of aspects of the use of test results, including process issues related to the communication of the test results in a clear, timely, useful, and confidential score report, as well as issues related to the consequences of the resulting decisions themselves (Bachman and Palmer, 2010, Ch. 9). For FSI, the last point would involve looking at outcomes related to the effectiveness of Foreign Service officers in carrying out language tasks and changes in the training program to develop language proficiency.

PROFESSIONAL STANDARDS FOR WORKPLACE TESTING

As suggested by the four example claims discussed above, evaluating the extent to which test scores can validly be interpreted and used as intended involves multiple investigations. For a high-stakes assessment program, investigating and establishing the validity of claims and documenting this process is a critical and ongoing feature of the overall program. Several professional organizations have articulated and published standards to guide the development and evaluation of assessment programs, focusing broadly on validity and the related issues of reliability, generalizability, and fairness.

In this section the committee offers an overview of the considerations raised by these professional standards. This overview does not detail the standards; rather, it highlights a set of best practices that are commonly discussed in the field and that work to ensure the validity of a test during its development, demonstrate its validity when it becomes operational, and guide the process of revision and renewal that is a necessary part of all ongoing testing programs. Some of these practices do not apply to all testing programs and all need to be shaped by a program's specific context, but the entire list provides a quick overview of key practices that testing programs should consider.

The committee reviewed nine sets of standards documents, paying specific attention to the guidelines most relevant to language assessment and assessment related to professional purposes. Two of these sets of standards focus specifically on language assessment: the International Language Testing Association *Guidelines for Practice*[2] and the European Association for

[2] See https://www.iltaonline.com/page/ILTAGuidelinesforPra.

Language Testing and Assessment Guidelines for Good Practice.[3] Three of the sets of standards address work-related assessment in the U.S. context: the *Standards for Educational and Psychological Testing* include a chapter devoted specifically to workplace testing and credentialing (American Educational Research Association et al., 2014); the *Uniform Guidelines on Employee Selection Procedures,*[4] which are an important source of information about fairness issues in employment-related decisions; and the *Principles for the Validation and Use of Personnel Selection Procedures* (Society for Industrial and Organizational Psychology, 2018), which provide instructions for conducting validation studies for job-related tests. Two of the sets of standards are from the International Test Commission and focus on issues related to international assessment: one related to the competencies needed by test users (International Test Commission, 2001) and the other related to assessment in settings that are linguistically or culturally diverse (International Test Commission, 2018). Finally, two sets of standards address job-related assessment issues in the context of accreditation: the *National Commission for Certifying Agencies Standards for the Accreditation of Certification Programs* (Institute for Credentialing Excellence, 2016) and the *American National Standards Institute* standards for personnel certification programs.[5]

These different standards documents lay out guidelines for identifying competencies to be assessed; developing the assessment and specific assessment tasks; field-testing assessment tasks; administering assessment tasks and scoring responses; setting the passing standard; and evaluating the reliability of the scores, the validity of interpretations based on the assessment results, and the fairness of the interpretations and uses of these results. Although the standards articulated in each document are tailored to different contexts, they address a number of common points.

The committee identified 11 best practices that are relevant for all high-stakes testing programs.

- **Practice 1: Work/Job Analyses.** In the context of employment settings, test developers should conduct a work/job analysis regularly (typically every 3–5 years) to document the work and worker-related requirements to which test content has been linked. For FSI, this practice relates to the development of the understanding of the target language use domain for tests of language proficiency.
- **Practice 2: Content-Related Validity Evidence.** Test developers should maintain documentation of the procedures used to deter-

[3] See http://www.ealta.eu.org/guidelines.htm.

[4] See http://www.uniformguidelines.com/uniform-guidelines.html.

[5] See https://www.ansi.org/accreditation/credentialing/personnel-certification.

mine the content (knowledge, skills, abilities, and other characteristics) to be covered by the test and the formats for test tasks. Documentation should include procedures for developing and field-testing items and for determining those items acceptable for operational use. Documentation should include the qualifications of those involved in this process.

- **Practice 3: Test Administration.** Test developers should provide detailed instructions for administering the test under standard conditions and for providing testing accommodations to test takers who need them. Instructions should also include testing policies and guidance for maintaining the confidentiality and security of test content.
- **Practice 4: Scoring Processes.** Test developers should maintain documentation about the scoring rubric for the test and the procedures used for scoring performances. This documentation includes procedures related to rater selection, training, recalibration, retraining, and monitoring and evaluation, as well as disagreement resolution, analyses of inter-rater reliability and agreement, and rater effects.
- **Practice 5: Cut-Score Setting.** If performance or passing levels are defined with respect to the score scale, test developers should maintain documentation about the process by which the cut scores defining the performance or passing levels were set. This documentation should include details about the qualifications of panelists used to set cut scores, the instructions they were given, their levels of agreement, and the feedback provided to them in score reports.
- **Practice 6: Psychometric and Other Technical Qualities.** Test developers should specify a program of research to evaluate the technical qualities of the test and maintain documentation on the progress and results of these studies. This documentation should include procedures for estimating reliability and decision consistency, evaluating rater agreement for the scoring of constructed-response questions, evaluating differential item functioning, if possible, and comparing performance by race, ethnicity, age, and gender. This research should also include other relevant technical analyses, such as cognitive studies to understand how test takers process the test questions and domain analyses to understand the relative weight given to different aspects of the target language use.
- **Practice 7: Fairness.** Test developers should maintain documentation about efforts to ensure that scores obtained under the specified test administrations procedures have the same meaning for all population groups in the intended testing population and that test takers with comparable proficiency receive comparable scores.

- **Practice 8: Score Reporting.** Test developers should maintain documentation about score report design, contents, and implementation.
- **Practice 9: Purposes and Uses of the Test.** Test developers should have written documentation of the needs, objectives, uses, and stakeholders for which a high-stakes testing program exists.
- **Practice 10: Explicit statement of the Validity Argument.** The key elements of a validity argument/framework should be available in an up-to-date technical report that is public. Organizations with effective high-stakes testing programs develop a good documentation culture over time.
- **Practice 11: Information for Test Takers.** Test developers should maintain up-to-date test-taker familiarization guides to reduce differences across test takers in familiarity with the test format.

These best practices provide a concrete guide for the different aspects of a testing program that should be evaluated to help ensure and establish the overall validity of the program's test results for its intended uses. In almost all cases, these considerations are relevant to FSI's testing program.

7

Balancing Evaluation and the Implementation of New Approaches

Throughout this report, the committee notes a number of specific options related to language assessment that FSI may want to explore, given the research literature on language assessment. The committee was specifically asked by FSI not to provide recommendations about FSI's language assessment program since the agency is responsible for determining its own program. Thus, in this concluding chapter the committee addresses the basic choice about the balance between evaluation and the implementation of new approaches that are relevant to FSI in determining how to proceed. Decisions about how to set this balance will influence all aspects of the development of FSI's assessment program.

BASIC CONSIDERATIONS

At the heart of the FSI's choice 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 could be changed in light of a principled approach to assessment, and (2) beginning the implementation of new approaches. Evaluation and implementation are both necessary: evaluation of the current program without implementation of new approaches to bring improvements will have no effect, and implementation of new approaches without a full evaluation of the current test could be very harmful to the current program. However, given limited time and resources, it is important to decide the relative attention to give to each.

Through this report, the committee addresses both evaluation, through the presentation of a principled approach to assessment, and implementation, through the presentation of new approaches to assessment. For evaluation, the report stresses the importance of understanding the target language use domain as a foundation for both the design and the validation of a testing program, and it briefly describes the available techniques for developing that understanding (Chapter 3). Furthermore, the report addresses the role the understanding of language plays in undergirding a testing program and the importance of understanding its sociocultural and institutional contexts. The report also stresses the importance of evaluating the validity of the uses of the test in light of the target language use domain and key details related to the test (Chapter 6). For implementation, the report suggests possible changes to the current test that might reflect identified goals for strengthening the current test, on the basis of what FSI currently knows about the strengths and weaknesses of the test or information that would result from an evaluation of the test (Chapter 4). These arguments and discussions raise the essential question about whether to emphasize—at this time—further evaluation to better understand the test and how well it is working or initial steps toward implementing plausible changes.

For FSI's decision about the relative attention to give to each, it will be important to consider how well FSI's current language assessment practices address the language proficiency needed by Foreign Service officers. The committee's limited understanding of the language proficiency needed by Foreign Service officers and the current language assessment suggests that there are certainly points of commonality. FSI's assessment is clearly different from a language assessment that might be used in other settings—such as certifying the language abilities of medical professionals or admitting graduate students to a course of study—and the distinctive aspects of FSI's assessment appear to reflect the language tasks of Foreign Service officers. However, the committee had insufficient evidence about the nature of the language proficiency needed and the alignment of FSI's assessment to those abilities to draw any conclusions about how close the alignment is. The committee's discussion of some possible changes to the current test highlights a number of ways that the coverage of the language proficiency of Foreign Service officers may be limited in capturing all important aspects of their language-related tasks, but the committee has no information about the relative importance of these omitted aspects of Foreign Service tasks.

One of the key issues for FSI to consider is whether it has sufficient information to draw firm conclusions about the degree of alignment between the aspects of language proficiency measured by the test and the aspects that affect the performance of key Foreign Service tasks. If the available information is not sufficient to draw firm conclusions, then obtaining better information about the alignment is particularly important. However, if there is already good information about the degree of alignment, then that information can help guide the consideration of changes to the current language assessment program.

The alignment between the language proficiency demonstrated by the current test and the language proficiency needed by Foreign Service officers is only one example of a key piece of evidence needed by a language assessment program to consider possible changes; it is captured in the first example validity claim discussed in Chapter 6. The other example validity claims discussed in that chapter suggest other instances of the tradeoff between evaluation and the implementation of new approaches. What is already known about the scoring process, the interpretation of the scores, and the relative benefits of the use of the scores? In each case, there could be very limited information, suggesting the importance of evaluation to improve understanding, or there could already be sufficient information to suggest that the test should be strengthened in some particular way or that there are no clear weaknesses.

One way to find a good balance between an evaluation of the current test and beginning implementation of new approaches to assessment is to consider the examples of validity evidence discussed in Chapter 6 and the best practices for testing programs recommended by the professional standards. For example:

- Does the FSI testing program have evidence related to the four example comparisons (pages 71–74)?
- Does the program incorporate the best practices recommended by the professional standards (pages 76–78)?

If the answer to either of these questions is “no,” then it makes sense to place more weight on the evaluation side, that is to first gather evidence to better understand how the current program is working. If the answer to these questions is “yes,” then there is probably already sufficient information to suggest particular ways that the test could be strengthened.

SOME CONSIDERATIONS ON THE EVALUATION SIDE

To the extent that FSI chooses to emphasize the evaluation side of the evaluation-implementation tradeoff, there are a number of important considerations. The discussions in Chapters 3 and 6 point toward a number of concrete questions that FSI could usefully further investigate related to the target language use domain and the different validity claims related to the current test. In addition, the possible changes discussed in Chapter 4 could each become a topic of evaluation, as an early step toward implementation. The exploration of new testing approaches on an experimental basis allows a testing program to better understand the tradeoffs of a change before any major decision to implement those approaches for an entire assessment program.

Beyond the specific evaluation questions themselves, there are questions about the institutional structure that supports evaluation research at FSI and provides an environment that fosters continuous improvement. Many assessment programs incorporate regular input from researchers into the operation of their program. This input can include two different elements. First, technical advisory groups are often used to provide an assessment program with regular opportunities for discussion of technical options with outside researchers who become familiar with the program's context and constraints during their service as advisors. Second, assessment programs also sometimes provide opportunities for researchers to work in-house as visiting researchers or interns to conduct research related to the program, such as conducting validity studies. Both of these routes allow assessment programs to receive new ideas from experts who come to understand the testing program and can provide tailored, useful advice. It is likely that there are constraints related to privacy and international security issues that could limit sharing data and publishing research on FSI outcomes, but it is possible that these constraints can be addressed with techniques to anonymize and share limited data for research. There also are costs associated with these activities, but many ongoing testing programs decide that these costs are outweighed by the long-term benefits of receiving regular input from outside researchers.

SOME CONSIDERATIONS ON THE IMPLEMENTATION SIDE

There are two salient constraints in the FSI testing program that are likely to strongly influence the consideration of possible new approaches. In the context of FSI's current testing program, these constraints appear to be fixed. However, it is worth considering the possibility that these constraints may be more flexible than currently presumed.

The first constraint relates to the policy that all languages should be assessed using the same approach. The fairness concerns that provide the foundation for this policy are understandable, but the comparability of results from the testing process is what actually matters for fairness, not an identical testing procedure. If the fairness issue can be addressed, it may be possible to consider using different testing approaches across languages.

It is worth noting that FSI's current assessment program already involves some limited variation in assessment procedures. The most prominent of these variations is the possibility for a test taker to interact with the evaluators over the phone or with a video conference rather than in person. In addition, in cases where only one assessor is available in a particular language, the assessor used for the test can be the test-taker's instructor. The general finding in the literature is that both of these variations can have an effect on assessment outcomes. However, in the context of FSI's assessment program, these variations are accepted as providing scores that are sufficiently comparable to those provided in a standard in-person assessment with an assessor who does not already know the test taker.

The reason to consider using different approaches to assess different languages is the practical implications of the number of test takers. Some assessment techniques—such as technology-mediated approaches—have relatively high development costs but relatively low administration costs per person. Thus, such techniques may be cost-effective only for relatively high-volume tests. In considering possible new language testing approaches, FSI needs to decide whether the practical limitations that might prevent the use of some approaches for the low-frequency languages should automatically disqualify their use for all languages.

The second constraint relates to the role of the Interagency Language Roundtable (ILR) framework. As FSI considers possible changes to its language assessments, it may want to consider options—such as the use of multiple measures—that may be awkward to score directly in terms of the ILR framework. However, the use of the ILR framework for coordination of personnel policies across government agencies does not need to be interpreted as a constraint requiring the use of ILR skill-level descriptions for all aspects of FSI scoring. As detailed in Chapter 5, whatever assessment approaches may be developed can always be mapped to the ILR framework for the purposes of final scoring and the determination of language proficiency.

The committee appreciates that FSI faces complicated choices about possible changes to its language proficiency testing, and the agency's interest in exploring the many aspects of modern language testing is commendable. The committee hopes that this report's discussion of the research in the field contributes to FSI's forward-looking decision process.

References

ACT, Inc. (2012). *Foreign Service Officer Test: Study Guide, 5th Edition*. Iowa City: ACT, Inc.

Alderson, C. (2000). *Assessing Reading*. Cambridge, UK: Cambridge University Press.

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

Bachman, L.F. (1990). *Fundamental Considerations in Language Testing*. Oxford, UK: Oxford University Press.

Bachman, L.F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly 2*:1–34.

Bachman, L.F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. Turner, and C. Doe (Eds.), *Language Testing Reconsidered* (pp. 41–71). Ottawa, Canada: University of Ottawa Press.

Bachman, L.F., and A. Palmer. (1996). *Language Assessment in Practice: Designing and Documenting Useful Language Tests*. Oxford, UK: Oxford University Press.

Bachman, L.F., and A. Palmer. (2010). *Language Assessment in Practice*. Oxford, UK: Oxford University Press.

Baker, B., and A. Hope. (2019). Incorporating translanguaging in language assessment: The case of a test for university professors. *Language Assessment Quarterly 16*(4-5):408–425.

Barnett, E.A, P. Bergman, E. Kopko, V. Reddy, C.R. Belfield, and S. Roy. (2018). *Multiple Measures Placement Using Data Analytics: An Implementation and Early Impacts Report*. Center for the Analysis of Postsecondary Readiness. Available: https://ccrc.tc.columbia.edu/publications/multiple-measures-placement-using-data-analytics.html.

Bernstein, J.C. (2013). Computer scoring of spoken responses. In C.A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 1–7). New York: Blackwell Publishing.

Brannick, M.T., K. Pearlman, and J.I. Sanchez. (2017). Work analysis. In J.L. Farr and N.T. Tippins (Eds.), *Handbook of Employee Selection* (pp. 134–162). New York: Routledge.

Brindley, G. (1994). Task-centered assessment in language learning: The promise and the challenge. In N. Bird, P. Falvey, A.B.M. Tsui, D. Allison, and A. McNeill (Eds.), *Language and Learning* (pp. 73–94). Hong Kong: Institute of Language in Education, Hong Kong Department of Education.

Brown, J.D, and T. Hudson. (1998). The alternatives in language assessment. *TESOL Quarterly 32*(4):653–675. doi:10.2307/3587999.

Brown, J.D., T. Hudson, J. Norris, and W. Bonk. (2002). *An Investigation of Second Language Task-based Performance Assessments*. Honolulu, HI: University of Hawaii Press.

Buck, G. (2001). *Assessing Listening*. Cambridge, UK: Cambridge University Press.

Canagarajah, S. (2006). Changing communicative needs, revised assessment objectives: Testing English as an international language. *Language Assessment Quarterly 3*(3):229–242.

Canale, M. (1983). On some dimensions of language proficiency. In J. Oller (Ed.), *Issues in LT Research* (pp. 333–342). Rowley, MA: Newbury House.

Canale, M., and M. Swain. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics 1*(1):1–47.

Carroll, J.B. (1961). Fundamental considerations in testing English proficiency of foreign students. In *Testing the English Proficiency of Foreign Students* (pp. 30–40). Washington, DC: Center for Applied Linguistics.

Cenoz, J. (2013). Defining multilingualism. *Annual Review of Applied Linguistics 33*:3–18.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing 20*(4):369–383.

Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. In. L. Bachman and A.D. Cohen (Eds.), *Interfaces between Second Language Acquisition and Language Testing Research* (pp. 32–70). Cambridge, UK: Cambridge University Press.

Chen, L., K. Zechner, S.Y. Yoon, K. Evanini, X. Wang, A. Loukina, and B. Gyawali. (2018). Automated scoring of nonnative speech using the speech rater SM v. 5.0 engine. *ETS Research Report Series 18*(10):1–31. Available: https://doi.org/10.1002/ets2.12198.

Chester, M.D. (2005). Making valid and consistent inferences about school effectiveness from multiple measures. *Educational Measurement: Issues and Practice 24*(4):40–52.

Chun, C.W. (2008). Comments on "Evaluation of the usefulness of the Versant for English test: A response": The author responds. *Language Assessment Quarterly 5*(2):168–172. doi.org/10.1080/15434300801934751.

Ciechanowski, L., A. Przegalinska, M. Magnuski, and P. Gloor. (2019). In the shades of the uncanny valley: An experimental study of human-chatbot interaction. *Future Generation Computer Systems 92*:539–548.

Cizek, G.J. (Ed.). (2012). *Setting Performance Standards: Concepts, Methods, and Perspectives, 2nd Edition*. New York: Routledge.

Cizek, G.J., and M.B. Bunch. (2007). *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests*. Thousand Oaks, CA: Sage.

Clark, J.L.D. (1972). *Foreign Language Testing: Theory and Practice*. Philadelphia: Center for Curriculum Development.

Council of Europe. (2018). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment; Companion Volume with New Descriptors*. Available: https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989.

Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. *Language Assessment Quarterly 10*(1):1–8.

Cummins, P.W., and C. Davesne. (2009). Using electronic portfolios for second language assessment. *The Modern Language Journal 93*:848–867. doi.org/10.1111/j.1540-4781.2009.00977.x.

Cushing-Weigle, S.C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing 9*(1):27–55.

Davies, A. (2003). *The Native Speaker: Myth and Reality.* Tonawanda, NY: Multilingual Matters Ltd.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing 26(*3):367–396. Available: https://doi.org/10.1177/0265532209104667.

Dewaele, J.M. (2018). Why the dichotomy 'L1 versus LX user' is better than 'native versus non-native speaker. *Applied Linguistics 39*(2):236–240.

Dorsey, D.W. (2005). The portfolio as a multipurpose tool: Part 1–using the portfolio for leadership development. In R. Mueller-Hanson's and D. Dorsey's (Chairs), The Portfolio: An Innovative Approach to Assessment, Development, and Evaluation. Practitioner Forum conducted at the Twentieth Annual Conference of the Society for Industrial and Organizational Psychology, Los Angeles, California.

Douglas, D. (2000). *Assessing Languages for Specific Purposes.* Cambridge, UK: Cambridge University Press.

Douglas, K.M., and R.J. Mislevy. (2010). Estimating classification accuracy for complex decision rules based on multiple scores. *Journal of Educational and Behavioral Statistics 35*(3):280–306.

Ducasse, A.M., and A. Brown. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing 26*(3):423–443. doi.org/10.1177/0265532209104669.

East, M. (2015). Coming to terms with innovative high-stakes assessment practice: Teachers' viewpoints on assessment reform. *Language Testing 32*(1):101–120. doi.org/10.1177/0265532214544393.

East, M. (2016). *Assessing Foreign Language Students' Spoken Proficiency: Stakeholder Perspectives on Assessment Innovation.* New York: Springer.

East, M., and A. Scott. (2011). Assessing the foreign language proficiency of high school students in New Zealand: From the traditional to the innovative. *Language Assessment Quarterly 8*(2):179–189. doi.org/10.1080/15434303.2010.538779.

Ferrara, S., E. Lai, A. Reilly, and P.D. Nichols. (2017). Principled approaches to assessment design, development and implementation. In A.A. Rupp and J.P. Leighton (Eds.), *The Handbook of Cognition and Assessment, Frameworks, Methodologies and Applications* (pp. 41–74). West Sussex, UK: Wiley.

Figueras, N., and J. Noijons (Eds.). (2009). *Linking to the CEFR Levels: Research Perspectives*. Arnhem: CITO and EALTA. Available: http://www.coe.int/t/dg4/linguistic/Proceedings_CITO_EN.pdf.

Forscher, P.S., C.K. Lai, J. Axt, C.R. Ebersole, M. Herman, P.G. Devine, and B.A. Nosek. (2016, August 15; preprint). A Meta-Analysis of Procedures to Change Implicit Measures. doi.org/10.31234/osf.io/dv8tu.

Friedrich, P. (Ed.). (2016). *English for Diplomatic Purposes*. Bristol, UK: Multilingual Matters.

Fulcher, G., and F. Davidson (Eds.). (2012). *The Routledge Handbook of Language Testing*. New York: Routledge.

Gass, S., and M. Varonis. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning 34*(1):65–89.

Gatewood, R.D., H.S. Feild, and M.R. Barrick. (2015). *Human Resource Selection*. Mason, OH: South-Western, Cengage Learning.

Gebril, A., and L. Plakans. (2014). Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks. *Assessing Writing 21*:56–73.

Gorter, D., and J. Cenoz. (2017). Language education policy and multilingual assessment. *Language and Education 31*(3):231–248.

Green, A. (2014). *Exploring Language Assessment and Testing*. Routledge Introductions to Applied Linguistics. New York: Routledge.

Greenwald, A.G., D.E. McGhee, and J.L. Schwartz. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology 74*(6):1464–1480.

Griffin, P., B. McGraw, and E. Care (Eds.). (2012). *Assessment and Teaching of 21st Century Skills*. New York: Springer Science+Business, Media.

Hambleton, R.K., and M.J. Pitoniak. (2006). Setting performance standards. In R.L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 433–470). Westport, CT: Praeger.

Hambleton, R.K., and A.L. Zenisky. (2013). Reporting test scores in more meaningful ways: A research-based approach to score report design. In K.F. Geisinger, B.A. Bracken, J.F. Carlson, J.-I.C. Hansen, N.R. Kuncel, S.P. Reise, and M.C. Rodriguez (Eds.), *APA Handbooks in Psychology®. APA Handbook of Testing and Assessment in Psychology, Vol. 3. Testing and Assessment in School Psychology and Education* (pp. 479–494). American Psychological Association. Available: https://doi.org/10.1037/14049-023.

Hart-Gonzalez, L. (1994). *Raters and Scales in Oral Proficiency Testing: The FSI Experience*. Paper presented at the Annual Language Testing Research Colloquium, Washington, DC. Available: https://www.semanticscholar.org/paper/Raters-and-Scales-in-Oral-Proficiency-Testing%3A-The-Hart-Gonz%C3%A1lez/ff6a160d3dcc091f7d78a1db6a308572123333c7.

Herman, J.L., M. Gearhart, and E.L. Baker. (1993). Assessing writing portfolios: Issues in the validity and meaning of scores. *Educational Assessment 1*(3):201–224. Available: http://dx.doi.org/10.1207/s15326977ea0103_2.

Housen, A., F. Kuiken, and I. Vedder. (2012). *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*. Amsterdam: John Benjamins.

Hu, G. (2018). The challenges of world Englishes for assessing English proficiency. In E.L. Low and A. Pakir (Eds.), *World Englishes: Rethinking Paradigms* (pp. 78–95). New York: Routledge.

In'nami, Y., and R. Koizumi. (2011). Structural equation modeling in language testing and learning research: A review. *Language Assessment Quarterly 8*(3):250–276.

Institute for Credentialing Excellence. (2016). *National Commission for Certifying Agencies (NCCA) Standards for the Accreditation of Certification Programs*. Available: https://www.credentialingexcellence.org/ncca.

International Test Commission. (2001). International guidelines for test use. *International Journal of Testing 1*(2):93–114.

International Test Commission. (2018). *ITC Guidelines for the Large-scale Assessment of Linguistically and Culturally Diverse Populations*. Available: www.InTestCom.org.

Isaacs, T. (2008). Towards defining a valid assessment criterion of pronunciation proficiency in non-native English-speaking graduate students. *Canadian Modern Language Review 64*(4):555–580.

Isbell, D., and P. Winke. (2019). ACTFL Oral Proficiency Interview – computer (OPIc). *Language Testing 36*(3):467–477. doi.org/10.1177/0265532219828253.

Jenkins, J. (2006). Current perspectives on teaching world Englishes and English as a lingua franca. *TESOL Quarterly 40*(1):157–181. doi.org/10.2307/40264515.

Jonsson, A. (2014). Rubrics as a way of providing transparency in assessment. *Assessment & Evaluation in Higher Education 39*(7):840–852.

Kachru, B.B. (1996). The paradigms of marginality. *World Englishes 15*(3):241–255. doi.org/10.1111/j.1467-971X.1996.tb00112.x.

Kane, M.T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (pp. 17–64). Westport: American Council on Education/Praeger Publishers.

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

Kenyon, D.M., and V. Malabonga. (2001). Comparing examinee attitudes toward computer-assisted and other oral proficiency assessments. *Language Learning and Technology 5*(2):60–83.

Knoch, U., and S. Macqueen. (2020). *Assessing English for Professional Purposes*. Routledge.

Knoch, U., and W. Sitajalabhorn. (2013). A closer look at integrated writing tasks: Towards a more focused definition for assessment purposes. *Assessing Writing 18*(4):300–308.

Koretz, D. (1998). Large-scale portfolio assessments in the US: Evidence pertaining to the quality of measurement. *Assessment in Education 5*:309–334.

Koretz, D.M., and L.S. Hamilton. (2006) Testing for Accountability in K–12. In R. L. Brennan (Ed.), *Educational Measurement* (pp. 531–578). American Council on Education/Praeger Publishers.

Kress, G. (2010). *Multimodality: A Social Semiotic Approach to Contemporary Communication*. New York: Routledge.

Kunnan, A.J. (2018). *Evaluating Language Assessments*. New York: Routledge.

Lado, R. (1961). *Language Testing*. New York: McGraw-Hill.

Lazaraton, A., and L. Davis. (2008). A microanalytic perspective on discourse, proficiency, and identity in paired oral assessment. *Language Assessment Quarterly 5*(4):313–335. doi.org/10.1080/15434300802457513.

Lei, L., and D. Liu. (2019). Research trends in applied linguistics from 2005 to 2016: A bibliometric analysis and its implications. *Applied Linguistics 40*(3):540–561.

Lemke, J. (2002). Travels in hypermodality. *Visual Communication 1*:299–325. doi.org/10.1177/147035720200100303.

Levashina, J., C.J. Hartwell, F.P. Morgeson, and M.A. Campion. (2014). The structured employment interview: Narrative and quantitative review of the research literature. *Personnel Psychology 67*(1):241–293.

Levinson, S. (1983). *Pragmatics*. Cambridge, UK: Cambridge University Press.

Little, D. (2011). The common European framework of reference for languages, the European language portfolio, and language learning in higher education. *Language Learning in Higher Education 1*(1):1–21. doi.org/10.1515/cercles-2011-0001.

Livingston, S.A., and M.J. Zieky. (1982). Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service. Available: https://www.ets.org/Media/Research/pdf/passing_scores.pdf.

Long, M.H. (2015). *Second Language Acquisition and Task-based Language Teaching*. Malden, MA: Wiley Blackwell.

Long, M.H., and J.M. Norris. (2000). Task-based teaching and assessment. In M. Byram (Ed.), *Encyclopedia of Language Teaching* (pp. 597–603). London, UK: Routledge.

Luecht, R.M., T. Brumfield, and K. Breithaupt. (2006). A Testlet assembly design for adaptive multistage tests. *Applied Measurement in Education 19*(3):189–202. (Special edition on multistage testing.)

Luecht, R.M., and R.J. Nungester. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement 35*:229–249.

Luecht, R.M., and S.G. Sireci. (2011). *A Review of Models for Computer-Based Testing*. Research report 2011–12, College Board. Available: https://files.eric.ed.gov/fulltext/ED562580.pdf.

Luke, S.D., and A. Schwartz. (2007). Assessment and Accommodations, Evidence for Education, National Dissemination Center for Children with Disabilities. Available: https://successforkidswithhearingloss.com/beta/wp-content/uploads/2013/09/Assessment-Accommodations-NICYC.pdf.

Luoma, S. (2004). *Assessing Speaking*. Cambridge, UK: Cambridge University Press.

Malabonga, V., D.M. Kenyon, and H. Carpenter. (2005). Self-assessment, preparation and response time on a computerized oral proficiency test. *Language Testing 22*(1):59–92.

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing 26*(3):397–421. doi.org/10.1177/0265532209104668.

McNamara, T. (1996). *Measuring Second Language Performance*. New York: Longman.

Messick, S. (1989). *Validity*. In R.L. Linn (Ed.), *Educational Measurement, 3rd Ed.* (pp. 13–103). New York: American Council on Education/Macmillan.

Messick, S. (1996). Validity and washback in language testing. *Language Testing 13*:243–256.

Mislevy, R.J. (2018). *Sociocognitive Foundations of Educational Measurement*. New York/London: Routledge.

Mislevy, R.J., and G. Haertel. (2006). Implications for evidence centered design for educational assessment. *Educational Measurement: Issues and Practice 25*:6–20.

Mislevy, R.J., L.S. Steinberg, and R.G. Almond. (1999a). *Evidence-centered Assessment Design*. Princeton, NJ: Educational Testing Service.

Mislevy, R.J., L.S. Steinberg, and R.G. Almond. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives 1*(1):3–62. doi.org/10.1207/S15366359MEA0101_02.

Mislevy, R.J., L.S. Steinberg, F.J. Breyer, R.G. Almond, and L. Johnson. (1999b). A cognitive task analysis with implications for designing simulation-based performance assessment. *Computers in Human Behavior 15*(3-4):335–374.

MLA Ad Hoc Committee on Foreign Language. (2007). Foreign languages and higher education: New structures for a changed world. *Profession 12*:234–245.

National Council on Measurement in Education. (1995). *Code of professional responsibilities in educational measurement*. National Council of Measurement in Education. Washington, DC: Author.

National Research Council. (2001a). *Building a Workforce for the Information Economy*. Committee on Workforce Needs in Information Technology; Computer Science and Telecommunications Board; Board on Testing and Assessment; Board on Science, Technology, and Economic Policy; and Office of Scientific and Engineering Personnel. Washington, DC: National Academy Press.

National Research Council. (2001b). *Knowing What Students Know: The Science and Design of Educational Assessment*. Committee on the Foundations of Assessment. J. Pellegrino, N. Chudowsky, and R. Glaser (Eds.). Washington, DC: National Academy Press.

National Research Council. (2008). *Assessing Accomplished Teaching: Advanced-level Certification Programs*. Committee on Evaluation of Teacher Certification by the National Board for Professional Teaching Standards. M.D. Hakel, J.A. Koenig, and S.W. Elliott (Eds.). Washington, DC: The National Academies Press.

National Research Council. (2014). *Developing Assessments for the Next Generation Science Standards*. Committee on Developing Assessments of Science Proficiency in K–12. Board on Testing and Assessment and Board on Science Education, J.W. Pellegrino, M.R. Wilson, J.A. Koenig, and A.S. Beatty (Eds.). Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Newton, J., and E. Kusmierczyk. (2011). Teaching second languages for the workplace. *Annual Review of Applied Linguistics 31*:74–92. doi.org/10.1017/S0267190511000080.

Nordquist, R. (2020). *Definition and Examples of Language Varieties*. ThoughtCo. Available: thoughtco.com/language-variety-sociolinguistics-1691100.

Norris, J.M. (2016). Current uses for task-based language assessment. *Annual Review of Applied Linguistics 36*(13):230–244. doi.10.1017/S0267190516000027.

Norris, J.M., J.D. Brown, T.D. Hudson, and W. Bonk. (2002). Examinee abilities and task difficulty in task-based second language performance assessment. *Language Testing 19*(4):395–418. doi.org/10.1191/0265532202lt237oa.

Norton, J. (2005). The paired format in the Cambridge speaking tests. *ELT Journal 59*(4):287–297. doi.org/10.1093/elt/cci057.

O'Reilly, T., and J. Sabatini. (2013). *Reading for Understanding: How Performance Moderators and Scenarios Impact Assessment Design*. Research Report RR-13-31. Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2333-8504.2013.tb02338.x

Ockey, G.J. (2009). The effects of group members' personalities on a test taker's L2 group oral discussion test scores. *Language Testing 26*:161–186. doi.org/10.1177/0265532208101005.

Ockey, G.J., and E. Wagner. (2018a). An overview of interactive listening as part of the construct of interactive and integrated oral test tasks. In G. Ockey and E. Wagner (Eds.), *Assessing L2 Listening: Moving towards Authenticity* (pp. 179–192). Amsterdam and Philadelphia: John Benjamins.

Ockey, G.J., and E. Wagner. (2018b). *Assessing L2 Listening: Moving towards Authenticity*. Amsterdam and Philadelphia: John Benjamins.

Oh, S. (2019). Second language learners' use of writing resources in writing assessment. *Language Assessment Quarterly 17*(1):60–84. doi.org/10.1080/15434303.2019.1674854.

Organisation for Economic Co-operation and Development. (2003). *Education at a Glance: OECD Indicators 2003*. Available: http://www.oecd.org/site/worldforum/33703760.pdf.

Organisation for Economic Co-operation and Development. (2016). *PISA 2018: Draft Analytical Frameworks*. Paris, France: Author.

Organisation for Economic Co-operation and Development. (2018). *PISA 2015: Results in Focus*. Paris, France: Author.

Oswald, F.L., G. Mitchell, H. Blanton, J. Jaccard, and P.E. Tetlock. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology 105*(2):171–192.

Papageorgiou, S., R.J. Tannenbaum, B. Bridgeman, and Y. Cho. (2015). *The Association between TOEFL iBT® Test Scores and the Common European Framework of Reference (CEFR) Levels* (Research Memorandum No. RM-15-06). Princeton, NJ: Educational Testing Service. Available: https://www.ets.org/Media/Research/pdf/RM-15-06.pdf.

Pettis, J.C. (2014). *Portfolio-based Language Assessment (PBLA): Guide for Teachers and Programs*. Available: https://listn.tutela.ca/wp-content/uploads/PBLA_Guide_2014.pdf.

Plakans, L. (2009). Discourse synthesis in integrated second language writing performance. *Language Testing 26*(4):561–587.

Plakans, L. (2014). Written discourse. In A. Kunnan (Ed.), *The Companion to Language Assessment*. Somerset, NJ: Wiley and Sons.

Plakans, L., and A. Gebril. (2012). Using multiple tests in an integrated writing assessment: Source text use as a predictor of score. *Journal of Second Language Writing 22*:317–230.

Pulakos, E., and T. Kantrowitz. (2016). *Choosing Effective Talent Assessments to Strengthen Your Organization.* Society for Human Resource Management (SHRM) Foundation. Available: https://www.shrm.org/hr-today/trends-and-forecasting/special-reports-and-expert-views/documents/effective-talent-assessments.pdf.

Purpura, J.E. (2004). *Assessing Grammar.* Cambridge, UK: Cambridge University Press.

Purpura, J.E. (2016). Second and foreign language assessment. *The Modern Language Journal 100*(S1):190–208.

Purpura, J.E. (2017). *Assessing Meaning*. In E. Shohamy and L. Or (Eds.), *Encyclopedia of Language and Education, Vol. 7. Language Testing and Assessment*. New York: Springer International Publishing.

Purpura, J.E. (2019). Questioning the currency of second and foreign language proficiency exams as measures of 21st century competencies. Teachers College, Columbia University, The Arts and Humanities Distinguished Lecture Series, October 10, 2019. Available: https://vimeo.com/367018433.

Purpura, J.E., and C.E. Turner. (2018). *Using Learning-oriented Assessment in Test Development.* Auckland, New Zealand: Language Testing Research Colloquium.

Purpura, J.E., and J.W. Dakin. (2020). Assessment of the linguistic resources of communication. In C. Chapelle (Ed.), *The Concise Encyclopedia of Applied Linguistics: Assessment and Evaluation* (pp. 1–10). Oxford, UK: Wiley.

Qian, D.D. (2009). Comparing direct and semi-direct modes for speaking assessment: Affective effects on test takers. *Language Assessment Quarterly 6*(2):113–125.

Quaid, E. (2018). Output register parallelism in an identical direct and semi-direct speaking test: S case study. *International Journal of Computer-assisted Language Learning and Teaching 8*(2):75–91.

Ramanarayanan, V., K. Kvanini, and E. Tsuprun. (2020). Beyond monologues: Automated processing of conversational speech (pp. 176–191). In K. Zechner and K. Evanini (Eds.), *Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech*. New York: Routledge.

Read, J. (2000). *Assessing Vocabulary.* Cambridge, UK: Cambridge University Press.

Robinson, P. (2001). Task complexity, cognitive resources, and syllabus design: A triadic framework for examining task influences on SLA. In P. Robinson (Ed.), *Cognition and Second Language Instruction* (pp. 287–318). Cambridge, UK: Cambridge University Press.

Roever, C., and G. Kasper. (2018). Speaking in turns and sequences: Interactional competence as a target construct in testing speaking. *Language Testing 35*(3):331–355. doi.org/10.1177/0265532218758128.

Sackett, P.R., P.T. Walmsley, A.J. Koch, A.S. Beatty, and N.R. Kuncel. (2016). Predictor content matters for knowledge testing: Evidence supporting content validation. *Human Performance 29(*1):54–71.

Schaeffer, G.A., B. Bridgeman, M.L. Golub-Smith, C. Lewis, M.T. Potenza, and M. Steffen. (1998). Comparability of paper-and-pencil and computer adaptive test scores on the GRE® general test. *ETS Research Report Series* (2):i–25. Available: https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.1998.tb01787.x.

Schissel, J.L., C. Leung, and M. Chalhoub-Deville. (2019). The Construct of Multilingualism in Language Testing. *Language Assessment Quarterly 16*(4-5):373–378. doi.org/10.1080/15434303.2019.1680679.

Seidlhofer, B. (2009). Common ground and different realities: World Englishes and English as a lingua franca. *World Englishes 28*(2):236–245. doi.org/10.1111/j.1467-971X.2009.01592.x.

Shavelson, R.J., and N.M. Webb. (1991). *Generalizability Theory: A Primer*. Newbury Park, CA: Sage.

Shohamy, E. (2011). Assessing multilingual competencies: Adopting construct valid assessment policies. *The Modern Language Journal 95*:418–429.

Skehan, P. (1998). *A Cognitive Approach to Language Teaching*. Oxford, UK: Oxford University Press.

Skehan, P. (2003). Task-based instruction. *Language Teaching 36*(1):1–12.

So, Y., M.K. Wolf, M.C. Hauck, P. Mollaun, P. Rybinski, D. Tumposky, and L. Wang. (2015). *TOEFL Junior Design Framework* (TOEFL). Young Students Research Report No. TOEFL Jr-02). Princeton, NJ: Educational Testing Service.

Society for Industrial and Organizational Psychology. (2018). *Principles for the Validation and Use of Personnel Selection Procedures, 5th edition*. Bowling Green, OH: Author.

Tannenbaum, R.J., and E.C. Wylie. (2008). *Linking English Language Test Scores onto the Common European Framework of Reference: An Application of Standard-setting Methodology*. TOEFL iBT Research Report RR-08-34. Princeton, NJ: Educational Testing Service. dx.doi.org/10.1002/j.2333-8504.2008.tb02120.x.

Taylor, L. (Ed.). (2011). *Examining Speaking: Research and Practice in Assessing Second Language Speaking*. Cambridge, UK: Cambridge University Press.

Taylor, L., and P. Falvery (Eds.). (2007). *IELTS Collected Papers: Research in Speaking and Writing Assessment.* Cambridge, UK: Cambridge University Press.

Tillema, H., M. Leenknecht, and M. Segers. (2011). Assessing assessment quality: Criteria for quality assurance in design of (peer) assessment for learning–a review of research studies. *Studies in Educational Evaluation 37*(1):25–34.

Turner, C.E., and J.E. Purpura. (2016). Learning-oriented assessment in second and foreign language classrooms. In D. Tsagari and J. Baneerjee (Eds.), *Handbook of Second Language Assessment* (pp. 255–272). Boston, MA: De Gruyter, Inc.

Van der Linden, W.J., and C.A. Glas (Eds.). (2000). *Computerized Adaptive Testing: Theory and Practice*. Dordrecht, The Netherlands: Kluwer Academic.

Van Moere, A. (2013). *Raters and Ratings*. Wiley Online Library.

VanPatten, B., J. Williams, and S. Rott. (2004). *Form-meaning Connections in Second Language Acquisition*. In B. VanPatten, J. Williams, S. Rott, and M. Overstreet (Eds.), *Form-meaning Connections in Second Language Acquisition* (pp. 1–26). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Wagner, E. (2018). A comparison of l2 listening performance on tests with scripted or authenticated spoken texts. In G. Ockey and E. Wagner (Eds.), *Assessing L2 Listening: Moving towards Authenticity* (pp. 29–44). Amsterdam and Philadelphia: John Benjamins.

Wainer, H., N.J. Dorans, R. Flaugher, B.F. Green, and R.J. Mislevy. (2000). *Computerized Adaptive Testing: A Primer*. Oxfordshire, UK, and Philadelphia: Routledge.

Wang, Z., K. Zechner, and Y. Sun. (2018). Monitoring performance of human and automated scores for spoken responses. *Language Testing 35*(1):101–120. doi.org/10.1177/0265532216679451.

Weigle, S.C. (1998). Using FACETS to model rater training effects. *Language Testing 15*(2):263–287.

Weigle, S.C. (2002). *Assessing Writing*. Cambridge, UK: Cambridge University Press.

Winke, P. (2013). The effectiveness of interactive group oral for placement testing. In K. McDonough and A. Mackey (Eds.), *Second Language Interaction in Diverse Educational Contexts* (pp. 247–268). New York: John Benjamins.

Wolfram, W., C. Temple Adger, and D. Christian. (1999). *Dialects in Schools and Communities*. Mahwah, NJ: Erlbaum

Yamamoto, K., L. Khorramdel, and H.J. Shin. (2018). Introducing multistage adaptive testing into international large-scale assessments designs using the example of PIAAC. *Psychological Test and Assessment Modeling 60*(3):347–368.

Yan, D., A.A. von Davier, and C. Lewis (Eds.). (2014). *Computerized Multistage Testing: Theory and Applications*. Boca Raton, FL: CRC Press.

Zapata-Rivera, D. (Ed.). (2018). *Score Reporting Research and Applications*. New York / Oxon, UK: Routledge.

Zenisky, A., R.K. Hambleton, and R.M. Luecht. (2010). Multistage testing: Issues, designs, and research in W.J. van der Linden and C.A.W. Glas (Eds.), *Elements of Adaptive Testing, Statistics for Social and Behavioral Sciences*. Available: https://link.springer.com/content/pdf/10.1007%2F978-0-387-85461-8.pdf.

Zhang, Y., and C. Elder. (2009). Measuring the speaking proficiency of advanced EFL learners in China: The CET-SET solution. *Language Assessment Quarterly* 6(4):298–314. doi.org/10.1080/15434300902990967.

Zuengler, J., and E. Miller. (2006). Cognitive and sociocultural perspectives: Two parallel SLA worlds? *TESOL Quarterly 40*(1):35–58.

Appendix

Biographical Sketches of Committee Members and Staff

Dorry M. Kenyon (*Chair*) is senior fellow for assessment at the Center for Applied Linguistics (CAL) and acting director of CAL's collaborative assessment activities with WIDA, a consortium of state departments of education dedicated to supporting English-language learners in K–12 contexts. He also serves as chair of the five-member expert panel of the U.S. Defense Language Testing and Assessment Project. His work over 30 years has covered all aspects of designing, developing, validating, and operationalizing second and foreign language assessments through many large projects at state, national, and international levels. Previously, he taught German and English as a foreign or second language in the United States and abroad. He has a B.A. in German and economics from Bowdoin College, an M.T.S. in theology from Gordon-Conwell Theological Seminary, an M.A. in teaching English as a foreign language from the American University in Cairo, and a Ph.D. in measurement, applied statistics, and evaluation from the University of Maryland.

David Dorsey is vice president and director of the Business Development Division at the Human Resources Research Organization, with more than 25 years of experience as a human capital consultant, researcher, and manager. Previously, he worked in the U.S. government in the area of defense and intelligence and as a vice president at Personnel Decisions Research Institutes. His research and development work has been in the areas of assessing critical skills in national security settings (e.g., cyber, foreign language), understanding adaptive performance, innovating performance management, using modeling and simulation technologies for learning, understanding career paths, and building corporate-level data science platforms and communities. He is the recipient of two major research awards and an award for being a top leader in government. He is an elected fellow of the Society for Industrial and Organizational Psychology (Division 14 of the American Psychological Association). He has a B.A. in psychology, an M.A. in industrial/organizational psychology, and a Ph.D. in industrial/organizational psychology, all from the University of South Florida.

Stuart W. Elliott (*Study Director*) is a scholar in the Division of Behavioral and Social Sciences and Education of the National Academies of Sciences, Engineering, and Medicine. Previously at the National Academies, as the director of the Board on Testing and Assessment, he led numerous studies on educational tests and indicators, assessment of science and 21st-century skills, applications of information technology, and occupational preparation and certification. He recently spent 3 years at the Organisation for Economic Co-operation and Development (OECD) working with PIAAC, the OECD's test of adult skills, resulting in a 2017 report, *Computers and the Future of Skill Demand*. He has a B.A. in economics from Columbia University and a Ph.D. in economics from the Massachusetts Institute of Technology, and he received postdoctoral training in cognitive psychology at Carnegie Mellon University.

Judith Koenig (*Senior Program Officer*) is on the staff of the Committee on National Statistics of the National Academies of Science, Engineering, and Medicine, where she directs measurement-related studies designed to inform education policy. Her work has included studies on the National Assessment of Educational Progress; teacher licensure and advanced-level certification; inclusion of special-needs students and English-language learners in assessment programs; setting standards for the National Assessment of Adult Literacy; assessing 21st-century skills; and using value-added methods for evaluating schools and teachers. Previously, she worked at the Association of American Medical Colleges and as a special education teacher and diagnostician. She has a B.A. in elementary and special education from Michigan State University, an M.A. in psychology from George Mason University, and a Ph.D. in educational measurement, statistics, and evaluation from the University of Maryland.

Lorena Llosa is associate professor of education in the Steinhardt School of Culture, Education, and Human Development at New York University. Her research focuses on second language teaching, learning, and assessment. Her work has addressed standards-based classroom assessments of language proficiency, assessment of academic writing, placement testing of U.S.-educated language minority students in community colleges, and the integration of language and content in instruction and assessment. Her research has appeared in a wide range of scholarly journals on all aspects of language testing and measurement, and she is associate editor of the *American Educational Research Journal*. She currently serves on the Committee of Examiners of the Test of English as a Foreign Language (TOEFL) at the Educational Testing Service. She has a B.A. in English and Spanish from Santa Clara University, an M.A. in teaching English as a second language from the University of California, Los Angeles, and a Ph.D. in applied linguistics with a specialization in language testing from the University of California, Los Angeles.

Robert J. Mislevy is Frederic M. Lord chair in measurement and statistics at the Educational Testing Service. His research interests center on applying recent developments in technology, statistical methodology, and cognitive research to practical problems in educational assessment and has covered Bayesian networks in educational assessment, Bayesian psychometric modeling, and the sociocognitive foundations of educational measurement. He is a recipient of the Lindquist Award for career contributions from the American Educational Research Association, the Samuel J. Messick Memorial Lecture Award from the International Language Testing Association, and a Career Contributions Award and four annual awards for technical contributions from the National Council on Measurement in Education. He served on the U.S. Defense Language Testing Advisory Panel, and he was a primary author of the final report of the National Assessment Governing Board's design feasibility team. He has a B.S. and an M.S. in mathematics from Northern Illinois University and a Ph.D. in research methodology from the University of Chicago.

Natalie Nielsen (*Consultant*) is an independent research and evaluation consultant whose work focuses on improving opportunities and outcomes for young people. Previously, she worked at the National Academies of Sciences, Engineering, and Medicine, first as a senior program officer for the Board on Science Education and then as the acting director of the Board on Testing and Assessment. She also previously served as the director of research at the Business-Higher Education Forum and as a senior researcher at SRI International. She has a B.S. in geology from the University of California, Davis, an M.S. in geological sciences from San Diego State University, and a Ph.D. in education from George Mason University.

Lia Plakans is a professor of education in foreign language and English as a second language at the University of Iowa and departmental executive officer for the university's Department of Teaching and Learning in the College of Education. Her research focuses on second language learning with particular emphasis on language assessment and literacy. Her work has explored assessments that integrate language skills, such as reading-into-writing tasks, to understand the underlying processes elicited, as well as the nature and scoring of performances that require integration. She is an associate editor for *Language Assessment Quarterly*, and she chairs the Committee of Examiners of the TOEFL at the Educational Testing Service. She has a B.A. in anthropology and psychology from the University of Iowa, an M.A. in teaching English as a second language/applied linguistics from Iowa State University, and a Ph.D. in foreign languages/English as a second language education from the University of Iowa.

James E. Purpura is professor of linguistics and education at Teachers College at Columbia University. His research interests include the assessment of second and foreign language grammar and pragmatics; second and foreign language test validation; and learning-oriented and scenario-based assessment in classroom and large-scale standardized assessment contexts. He currently serves on the expert panel of the U.S. Defense Language Testing and Assessment Project, and he is an expert member of the European Association of Language Testing and Assessment. He is coeditor of *Language Assessment Quarterly* and coeditor of two book series devoted to language assessment. He has served as president of the International Language Testing Association, on the Committee of Examiners for the TOEFL at the Educational Testing Service, and as a language testing consultant for numerous foreign countries. He has a B.A. in French language and literature from Marietta College, an M.A. in French linguistics from the University of Colorado, Boulder, and a Ph.D. in applied linguistics from the University of California, Los Angeles.

M. “Elvis” Wagner is associate professor of teaching English to speakers of other languages (TESOL) at Temple University. His research focuses on second language assessment, especially the assessment of second language listening ability, and it is informed by broad teaching experience in many different contexts, including as a Peace Corps volunteer in Poland. He has written and published widely on issues related to foreign and second-language teaching methodology, as well as teaching and testing of second language listening and oral communicative competence. He is a recipient of the award for best article in language testing by the International Language Testing Association for “Video Listening Tests: What Are They Measuring?,” which appeared in *Language Assessment Quarterly*. He has a B.A. in English, Spanish, and history from Texas Christian University, an M.A. in English from the University of Nebraska, and an M.A. and an M.Ed. in TESOL and an Ed.D. in applied linguistics from Teachers College, Columbia University.

Paula M. Winke is an associate professor of second language studies (SLS) at Michigan State University and codirector of the university's SLS Eye-tracking Lab. She teaches assessment for classroom and research purposes and individual differences in second language acquisition. Previously, she was a German instructor at the University of Minnesota and a test development manager at CAL in Washington, D.C. She also served as a Peace Corps volunteer in China and a Fulbright English-language instructor in Hungary. She is a past president of the Midwest Association of Language Testers, and she is a recipient of the distinguished researcher award of the TESOL International Association and the Article of the Year Award of the Computer-Assisted Language Instruction Consortium. She is editor of *Language Testing*. She has a B.A. in French and philosophy from the University of Wisconsin–Madison, an M.A. in linguistics from the University of Minnesota, and a Ph.D. in applied linguistics from Georgetown University.