Language Assessment in the US Government

Rachel L. Brooks

Abstract

The US government is one of the first and most influential language assessment organizations in the USA. With its foundation being the Interagency Language Roundtable (ILR) Skill Level Descriptions, the US government has developed and administered tests not only in proficiency skills (listening, reading, speaking, writing) but led the way in performance testing (translation, audio translation, and interpretation) and intercultural competence. The scope of testing in the US government is tens of thousands of tests administered annually in hundreds of languages. Important to the US government is its operational underpinnings; tests are developed and administered to meet the missions of the agencies. US government agency scores are used to make a wide range of high-stakes decisions that can impact not only the careers of the examinees but also the lives of people the world over. Tight deadlines and limited resources, as well as changing needs and complexities in language challenge government test developers. Research regarding US government language-testing examines issues such as the relationship between reading, writing and translation, rater characteristics, standard setting, and other topics meant to improve the quality of language testing. In recent years, the US government assessment programs have increased collaboration among agencies leading to additional resources and helping each agency better fulfill its mission.

Keywords

Government • Interagency Language Roundtable Skill Level Descriptions • Proficiency • Performance

R.L. Brooks (✉)
Federal Bureau of Investigation, Washington, DC, USA
e-mail: rachel.brooks@ic.fbi.gov.

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_4

Contents

Introduction

Government testing programs span different types of agencies such as diplomatic, military, clandestine, and investigative. These agencies are responsible for administering their own language-testing programs, but they share resources and information, often under the umbrella of the Interagency Language Roundtable (ILR). The ILR provides a venue for agencies to exchange ideas, hold symposia, and share research (Jones and Spolsky 1975; ILR 2016). The US government collectively conducts tens of thousands of tests annually in nearly 200 languages, covering all levels of proficiency. The government conducts tests in a range of skills: listening, reading, speaking, writing, translation (including document, audio, and summary), interpretation, and transcription.

US government language testing poses unique challenges. Testing is tailored to operational needs that shift based on world events, impacting the types of tests needed and requiring tight deadlines. US government language testing is high stakes because it determines whether government personnel have a reliable ability to perform the language tasks to support defense, diplomatic, national security, and law enforcement needs. Testing programs meet these challenges by developing new tests, as well as adapting and adopting available resources for assessments. Testing not only impacts examinees but also the agency mission and, consequently, the citizens the agencies serve.

Early Developments

In the US government, language learning and assessment programs have always focused on practical needs stemming from current events, such as wars, terrorist acts, and international events. Prior to the 1940s, the focus of language assessment was classroom assessments of reading proficiency. It was localized in each agency, with little interagency collaboration. The US involvement in World War II caused language training and testing efforts to increase significantly, leading to resource

sharing among agencies. Moreover, World War II shifted the focus of language learning from reading to listening and speaking. Radio transmissions became an integral part of wartime communication, leading to the need for foreign language intercepts. More and more soldiers were being deployed overseas, requiring conversational abilities. To meet these changes, Kaulfers (1944) outlined a methodology for aural and oral language evaluation, including rubrics and rating criteria. In 1949, the US Army released the first standardized tests of proficiency in reading, listening, writing, and grammar in 25 languages called the Army Language Tests (Pulliam and Ich 1968) based on Kaulfer's methodology.

The standardization of language testing also had an impact on language aptitude testing. Before World War II, US military language course placement was determined by a combination of measures, including IQ tests, general language aptitude tests, and tests of how well a person could speak a "first" language (Myron 1944). These tests were found to be ineffective measures of language aptitude once language training moved away from the translation method, leading to a formalized aptitude assessment (Petersen and Al-Haik 1976). One of two early aptitude tests was the Department of Defense's Defense Language Aptitude Test (DLAT). The Modern Language Aptitude Test (MLAT) followed the DLAT in 1959 and was widely used by agencies in both the USA and Canada. In 1976, the DLAT was revised, validated, and renamed the Defense Language Aptitude Battery (DLAB) (Petersen and Al-Haik 1976).

Before long, the Army Language Tests released in 1949 needed updating and in 1954 the Army Language School (now the Defense Language Institute Foreign Language Center (DLIFLC)) constructed the Defense Language Proficiency Tests (Pulliam and Ich 1968). Meanwhile, in 1952, the US Civil Service Commission was tasked with inventorying the language abilities of government employees across agencies, requiring standardized assessment criteria. Government personnel included native speakers, heritage speakers, and language learners, so a way to assess language proficiency regardless of how the language ability was attained was critical. The US government developed its own standardized criteria since no such criteria were found in academia (Herzog 2003; Jones and Spolsky 1975; Lowe 1985). The US Foreign Service Institute (FSI) of the Department of State came up with the first rating scale of functional language ability, with score levels 1–6. An independent testing office at FSI, established in 1958, extrapolated a format for reliable speaking testing from these criteria known as the "FSI test." In 1968, other US government agencies collaborated with FSI to develop and expand the criteria to cover speaking, listening, reading, and writing. This project resulted in the Interagency Language Roundtable (ILR) Skill Level Descriptions. Subsequently, federal government agencies worked to update and develop additional language tests based on the ILR. In particular, the FSI test was adapted for general proficiency use, expanding its breadth from the original FSI-focused scope, by a number of agencies and became known as the Oral Proficiency Interview (OPI) (Lowe 1988).

As the ILR Skill Level Descriptions were more broadly implemented across agencies, they received feedback and underwent revisions. The ILR scale adopted "plus" levels, which indicated language users with an ability that substantially

exceeded the base level, yet did not fully meet the next higher level. In 1985, the US Office of Personnel Management approved the ILR Skill Level Descriptions as the official criteria for evaluating the language proficiency of government personnel (Interagency Language Roundtable 1985). In the early twenty-first century, the ILR addressed the need to measure language in performance skills derived from operational language tasks such as translation, interpretation, transcription, and audio monitoring. The Translation and Interpretation Committee of the ILR joined with the Testing Committee to develop a set of performance skill level descriptions, including translation (2006), interpretation (2007), and audio translation (2011) (Brau 2013). Around the same time, discussions commenced on the importance of measuring the cultural knowledge and abilities used in communication between government personnel and native speakers overseas. To capture the progression of extralinguistic communication elements, the ILR developed the Skill Level Descriptions for Competence in Intercultural Communication (2012) (Interagency Language Roundtable 2016).

Major Contributions

Government Testing Criteria

The US government most often uses the ILR Skill Level Descriptions as their criteria for assessing language. The descriptions provide a common reference enabling organizations to have comparable expectations about general ability. They are an ordinal scale composed of six base levels from 0 to 5 with five plus levels from 0+ to 4+, totaling eleven ranges. They were developed by subject matter experts in language acquisition with experience in assessment representing the agencies that most frequently administer language testing (Lowe 1998). The ILR levels assume importance because most US government language tests use these scales as a reference. Therefore, they must be understood by all government stakeholders, including examinees, managers, training coordinators, etc. The descriptions do not provide comprehensive lists of abilities or linguistic functions and as such are subject to interpretation. The challenge in the production and use of the ILRs is that they must be general enough to meet the diverse needs of the agencies that use them, while being specific enough to control for reliable interpretation by the different organizations. The ILRs must meet the needs of the agencies that rely on them, which generally result in a lengthy development and approval process. Since the ILRs became the official language rating criteria for the US government, significant resources have been invested to develop and validate assessments based on them, including the Defense Language Proficiency Test (DLPT), the Oral Proficiency Interview (OPI), and the Verbatim Translation Exam (VTE). ILR-based tests look at a person's functional ability to perform linguistic job tasks specific to each agency and its validity lies in its ability to measure functional ability reliably. Agencies regularly conduct reliability checks from independent raters and have over the years

proved that the functional progression shown in the scales is accurate regardless of how the language was acquired (Brau 2013; Lowe 1988).

The ILR Skill Level Descriptions have importance outside the government context as well. They are the basis for the American Council on the Teaching of Foreign Languages (ACTFL) Guidelines, which were intentionally designed to be commensurate and derivative of the ILR. As such, the ACTFL Guidelines are at times used within the US government context, such as in the Peace Corps and the Department of Education. Additionally, the ILR Skill Level Descriptions heavily influenced the NATO STANAG (standardization agreement) 6001 language proficiency guidelines, which are used by foreign governments, including Canada and several European countries (Bureau for International Language Co-ordination 2016).

The framework of the ILR Skill Level Descriptions has important ramifications for developing and scoring language proficiency tests. First, the ILR Skill Level Descriptions are non-compensatory, that is, strength in one feature cannot compensate for weakness in another feature at a given level. For example, someone who can orally support opinions on societal-level topics using precise vocabulary (a level 3 skill) cannot be considered to have an overall level of 3 in speaking if there are persistent errors that interfere with comprehension, such as failure to distinguish singular and plural. Second, overall control of functions, a person's ability to accomplish particular language tasks, rather than total absence of errors or perfection of understanding are important (Brooks 2013).

Government Perspective

Since the major driving force behind government language testing is operational need, performance testing is essential. Within government contexts, the distinction between proficiency and performance testing has become significant. Proficiency testing refers to a holistic evaluation of a person's functional ability in the language. It is a general assessment that does not pay regard to how a language was acquired. The ILR scales for proficiency are the original four skills of listening, reading, speaking, and writing. When these first skill level descriptions were developed, testing focused on post language training exams. Assessing functional proficiency remains important because the government needs language generalists who have flexible language ability that can quickly meet needs. Government organizations highly value personnel who maintain high levels of general proficiency in a variety of skills.

In more recent years, it has become evident that testing of performance skills that require prerequisite proficiencies (i.e., translation which requires reading and writing proficiencies) is more practical than testing proficiency alone for government purposes. Performance tests, which measure a person's ability to perform a certain job, assess specific skills, such as translation, summarization, interpretation, and transcription all arise from operational tasks (Brau 2013; Child et al. 1991). Therefore, performance tests are a more practical and valid measure of the skills being used on

the job. Some agencies have worked to create performance tests since the late 1990s, but they are still only available in the top 30 or 40 tested languages. When performance tests are not available, testing programs have to rely on proficiency exams.

Impact of Agency Mission

The US government agency that is probably most well known for foreign language training and testing is the Department of State (DOS), which includes the Foreign Service Institute (FSI). The School of Language Studies at FSI is responsible for foreign language training of foreign service officers who interact with counterparts in US embassies. Its personnel have regular contact with counterparts from numerous international backgrounds, requiring high-level language skills, particularly in speaking. Diplomats need to converse with foreign counterparts, read foreign documents, and listen to broadcasts in other languages. Language Services at the Department of State has translators and interpreters that routinely perform specialized language tasks such as translation of international treaties and agreements and interpretation of negotiations and official addresses. Translators and interpreters are expected able to understand nuance, tone, implied meanings, and cultural references. Moreover, employees of diplomatic agencies serve as the face of their country in foreign lands; therefore, miscommunication could potentially lead to serious ramifications on international relations. Consequentially, diplomatic personnel typically endeavor to communicate effectively and appropriately as educated native speakers of the foreign language. Skills such as negotiation, persuasion, tact, and other influencing skills are expected to be mastered. Language testing emphasizes speaking but also reading and listening for officers and translation and interpretation for linguists at Language Services. The testing program is geared to high-level proficiency, ILR levels 3 and above as a goal.

Within the Department of Defense (DOD), foreign area officers, like diplomats, work in embassies and may need to negotiate and communicate agreements in security cooperation efforts between the USA and other countries. Primarily, however, defense organizations focus on giving military personnel the communicative skills they need to survive in foreign lands. They teach speaking and listening in routine or survival communications, such as gathering information from residents about local activities and performing security operations. Other personnel may monitor recorded or written communications from hostile groups. Although military personnel often do not need high levels of proficiency, the stakes are high. Inaccurate transfer of information could lead to loss of life or property. The majority of those trained and tested at the DOD take listening, reading, and speaking proficiency tests at ILR levels 3 and below.

In clandestine services, such as the Central Intelligence Agency (CIA) and National Security Agency (NSA), agents working undercover need to develop structural competence, vocabulary, and pronunciation that are parallel to those of native speakers. Additionally, they must acquire native speakers' cultural and

pragmatic skills, so as to be indistinguishable from them. Language errors have the potential to lead to loss of life or intelligence. Agents gather intelligence through audio intercepts, so listening skills are paramount. Listening comprehension tasks are complicated by the inability to ask for clarification and by poor recording quality. Additionally, a large number of language tasks require decoding vague, accented, slang, and veiled language. Language testers work to interpret how this type of task fits into the general rating scales and how to reliably assess listening in such contexts.

Investigative and law enforcement agencies, such as the Federal Bureau of Investigation (FBI) and the Drug Enforcement Agency (DEA), generally, serve both criminal and intelligence missions. Operational requirements demand that language personnel have both monitoring and translation abilities, with added legal requirements that govern the collection of and reporting on evidence and intelligence. Monitors overhear and then write analytical summaries of information relevant to investigations, which are often distinct from the main idea or supporting details of the audio. National privacy laws restrict material that can be monitored, so audio is truncated, causing additional listening challenges. Documents that are collected as evidence for investigations need to be translated so that the information is accessible to agents working on the related cases. Translation errors can lead to the dismissal of evidence admitted in court proceedings. As in government organizations, most interpretation assignments are informal and involve interviewing speakers of other languages. Investigative agencies also employ undercover agents who are high-level speakers of foreign languages. In all of these cases, single skill testing does not sufficiently measure language for the task, therefore performance testing of combined skills is increasing. Inaccuracies in court interpretations can result in unwarranted imprisonment or unprosecuted crimes. High levels of proficiency in speaking and listening do not necessarily result in high-quality interpretation. Therefore, most court systems test for interpretation skills directly rather than inferring them from the results of speaking proficiency tests.

In the USA, the Department of Education (DOE) oversees school curricula, initiatives, and assessments in all subject matters, including language. Educational institutions use language testing and their corresponding frameworks to measure the progress of student language learning. Education personnel referring to rating scales are generally interested in the lowest levels offered, as the majority of students will achieve results at these levels. Combined skills such as interpretation and translation are not taught except in specialized schools; therefore, educational agencies refer largely to the scales for the four primary skills using the ACTFL Guidelines. Often outcomes on these tests are used to measure student achievement and teacher performance.

In the US Peace Corps, humanitarian volunteers serve for one or two years in foreign countries teaching language or providing aid services. Most language learning that is done is in country and addresses survival needs rather than professional contexts; therefore, participants typically only achieve low levels of language proficiency. As in educational departments, service personnel may be tested via speaking proficiency tests to measure how much language learning was achieved. In other cases, such as the US National Language Service Corps, volunteers are

reserves. They are tested for general speaking proficiency so that, when a need arises, the organization knows which volunteers are most capable.

Increasingly, almost all aspects of government work are affected by foreign languages and all government agencies need some types of language users. Border officers need to conduct basic interviews, but they also need to be able to detect if a person is being dishonest. The Internal Revenue Service investigates and audits tax records and payments, requiring language personnel with reading skills to review records kept in foreign languages and writing skills to issue official letters in a language that the recipient can understand. Census workers conduct surveys in multiple languages to ensure accurate data collection and provide personnel capable of answering questions and conducting interviews with residents who have low levels of literacy to ensure accurate population statistics. All of the personnel that perform these duties need to undergo the appropriate level and type of language tests to ensure that their jobs are being done accurately, making language testing increasingly important to many government agencies.

Work in Progress

Research into language testing within the US government is largely focused on improving assessment to respond to changing needs in the agency. Language testers in the government produce, administer, and score tests to ensure continued quality results. A typical focus of research in the US government is quality assurance, validity, and efficiency, meaning how to produce results faster or using fewer resources.

In the mid- to late twentieth century, research paid attention to the impact of factors affecting the way the ILRs functioned. Higgs and Clifford (1982) investigated the proportions of rating factors (such as structures and vocabulary) contributing to ILR ratings. Child (1987) outlined the requirements for his ILR-based reading text typology. Lowe (2001) examined the wordings of the ILR Skill Level Descriptions at each level, examining best case, average case, and worst case statements and how these worked for rating in the four proficiency skills. These seminal works were accompanied by others that investigated the nature of the ILR scale and proficiency testing.

The US government's early use of only proficiency exams was based on the fact that most early examinees were native speakers of English and that native speakers of English only need to be tested in receptive skills in the foreign language. Research by Lunde and Brau (2005, 2006) investigated the correlation initially between reading and translation abilities and later between writing and translation abilities. The research found no significant correlation between strong translation ability and strong ability in either reading or writing, leading to the conclusion that a separate skill, the ability to transfer language from one language to another, was needed beyond knowledge of the two languages to successfully translate. In 2015, this research was updated with a larger data set including more languages and the same conclusions were drawn (Brooks and Brau 2015). Consequentially, it is not

advisable to use reading and writing proficiency tests to predict translation ability; translation tests should be administered.

Government language testers utilize hundreds of human raters evaluating a large number of exams, so there is a logical interest in rater reliability and the effects of various rater characteristics, such as native speaker status, rater language proficiency, and rater first language. Rater characteristic research has benefited from studies done within the government context, as it often deals with language proficiencies higher than those typically achieved through academic contexts and with more formalized, large-scale assessment. For example, Brooks (2013) showed how native speaker status has no significant impact on speaking test ratings but rater proficiency level does. The research supported the movement to remove references to the native speaker as a standard for assessment from testing documents and as a requirement for raters.

The importance of standard setting is recognized, and has been most widely used by DLIFLC for the DLPT. Beginning in 2009, the Department of Defense began standard-setting studies to set cut scores according to the ILR Skill Level Descriptions for the DLPT. A standard-setting study engages a panel of language experts who evaluate the item difficulty according to the ILR-SLDs and judge the likelihood of an examinee at a particular level of proficiency to succeed at each item (Impara and Plake 1997). The information provided by the judges, who also have access to pilot test data, is used in the calculation of cut scores for each ILR level. In addition, a larger-scale research effort is underway at the Department of Defense to isolate factors that affect difficulty of understanding audio material, beyond the factors referenced in the ILR Skill Level Descriptions. An initial study on the effect of the density of spoken texts on comprehension is in the planning stages.

The Testing and Assessment Expert Group (TAEG) is a focus group that operates under the Foreign Language Executive Committee (FLEXCOM) of the US Office of the Director of National Intelligence. It is made up of language-testing experts and representatives from various government agencies. TAEG conducted an unpublished interagency comparability study of speaking tests including three agencies and over 150 examinees conducted from 2009 to 2012. As a result of this study, there has been support for annual interagency comparability workshops where the four agencies with speaking test programs (CIA, DLI, FBI, and FSI) meet to review speaking tests and discuss protocol in an effort to better understand each other and norm to the ILR Skill Level Descriptions (Office for the Director of National Intelligence 2016).

Problems and Difficulties

US government language testers face a constant challenge. On the one hand, they are expected to provide assessments that meet operational demands in critical situations that may arise without warning, and at the same time, they maintain high standards of test validity and score reliability. This combined with the demand to administer thousands of tests annually in an increasing number of languages taxes government resources.

Fluctuating operational needs such as changes in language-related positions, responsibilities, and personnel often call for realignment of test batteries and passing scores or, in many cases, the development of an entirely new test. Often, there is not a large enough population of speakers of the tested language in order to trial the test thoroughly. Test developers must rely on modifying existing test instruments from within their agency or borrowing them from partner agencies. Production time frames by far less than needed for development and validation. Often test development deadlines must be met without additional funds or personnel. Developers rely on in-house technical personnel paired with translators from the field to produce the needed instrument.

The broad range of languages needed and classification of those languages and dialects pose challenges. The US government regularly has a need to communicate or process work in hundreds of languages, representing most language families. Acquiring, training, and evaluating personnel for so many languages pose challenges. Further, many languages have multiple variants or dialects and decisions need to be made as to whether or not it is appropriate to test them separately. Such decisions are often guided by considerations of mutual intelligibility and established recognition of the languages as separate and operational needs; all of these considerations may change with time. For example, Serbo-Croatian was once tested as a single language, but Serbian, Croatian, and Bosnian are now considered independent languages. These decisions are necessary but also costly.

Since the ILR Skill Level Descriptions are used across multiple languages, there are challenges in how to interpret language proficiency equivalently when languages function differently. Issues of diglossia and the acceptability of other "foreign" language features are of issue in language evaluation. Indian subcontinent languages such as Hindi, Punjabi, and Gujarati incorporate a lot of English, and it would at times be incorrect or inappropriate to use the Hindi/Punjabi/Gujarati word in certain contexts even when one exists. Moreover, creoles and patois often convert to other languages when certain proficiency levels are reached. For example, Haitian Creole becomes French for certain functions and contexts. When high-level language functions require shifting to another language, government agencies are challenged to decide whether the upper level functions can be supported by the test language and, therefore, whether or not an examinee can reach the highest level of the scale in that language (Brooks and Mackey 2009).

In Arabic dialects, for example, professional, sophisticated, or contextualized language tasks would never be conducted in the dialect, but rather in Modern Standard Arabic (MSA). It is for this reason that many US government agencies are shifting from testing Arabic dialects in isolation to testing the dialect combined with MSA, particularly in speaking exams. In 2010, the FBI began combining the tests, followed by FSI shortly thereafter. Combined Arabic testing is now being adopted by other agencies. MSA-only tests still exist to evaluate the language of personnel who have taken MSA training courses.

Government language evaluators are challenged to educate the test score users within the organization: the managers, the operational staff that need linguists, and the examinees themselves. Typically, test score users are not accustomed to the

nature of language or are not familiar with the ILR Skill Level Descriptions, leading to confusion, misunderstanding, and inappropriate score use. The indeterminate nature of language, with endless room for interpretation, can lead users to the conclusion that the language test scores are grossly subjective and therefore not accurate. Examinees often misinterpret their ratings' corresponding descriptions to mean the entirety of what a person can do, not the minimum threshold of that level. Likewise, untrained users can misinterpret what a score represents and assign an inappropriate operational task such as giving a translation task to an individual with a high speaking score. To combat this misuse of scores, many US government agencies now provide assessment literacy trainings to examinees and other stakeholders. The trainings are tailored to particular stakeholder audiences to help understand the nature of the ILR scales, how ratings are assigned and how they can be interpreted.

Future Directions

The focus for government language testing has historically been on producing a useful product that meets the immediate need. Although there have been guidelines for individual tests developed, there have not been set US government standards for quality of language tests or requirements for language-testing procedures; these standards have been left to the individual agencies. With the initiation of the newest generation of DLPTs in 2000, language-testing professionals were being hired by the DLIFLC to support the initiative. The professionalization effort advanced in 2009, when government language testers formed a subcommittee under the American Society for Tests and Materials (ASTM) to write a standard practice for ILR-based language proficiency testing. This standard practice was produced through collaboration between government personnel from many different agencies and private sector language-testing professionals (ASTM 2011).

There are two US government-based organizations that allow for collaboration among agencies with testing programs and needs: the Testing and Assessment Expert Group (TAEG) and the ILR Testing Committee. TAEG is a group formed under the Foreign Language Expert Group of the Office of the Director of National Intelligence. Its membership is composed entirely of government employees who are either language-testing experts or significant language-testing stakeholders. The committee meets monthly to share information and produce official recommendations and cross-agency initiatives. They catalog all the language-testing capacities of the agencies as well as the standards used for test development and quality assurance. Additionally, they have produced recommendations on quality translation assessment and research the comparability of test scores among agencies. Organizations like TAEG are essential to meeting operational needs, as many of the languages that suddenly become critical for an agency's mission are rarely used or assessed in the USA.

The ILR Testing Committee has long been a venue for collaboration and information sharing among government agencies. Its membership is composed not only

of government employees but also of members of academia and industry. The committee has taken on several projects to promote assessment literacy, including understanding the ILR and the development of self-assessment checklists to accompany the ILR Skill Level Descriptions (Interagency Language Roundtable 2016). The ILR Testing Committee has been involved in efforts to clarify and annotate the ILR Skill Level Descriptions for speaking, reading, and listening, to which end there have been several summits involving government and private sector language-testing professionals coming together to discuss the ILR Skill Level Descriptions and articulate a common interpretation of them.

Recent discussions within the TAEG and the ILR Testing Committee have led to a new initiative to revise the four original proficiency skill level descriptions for listening, reading, speaking, and writing. A subcommittee under the ILR Testing Committee has taken on the task of revising the listening descriptions first (Interagency Language Roundtable 2016). The goal of the revisions is not to change the core meaning of each level, which has been in use for over 30 years, but rather to update them, to remove references to antiquated technologies, integrate new modes of communication that have been introduced, clarify and expand upon some of the supporting statement, and remove controversial and difficult to identify concepts, such as the "native speaker" (Brooks 2013).

The top priority of US government assessment is ensuring that government language personnel are qualified to perform the mission of their agencies. US government agencies have a large number of challenges to overcome: developing appropriate language evaluations for an ever-increasing range of languages with minimal resources under strict time constraints for multiple skills, levels, and purposes, all while maintaining a high level of quality. The US government has been a leader in government language testing and has collaborated with government agencies of other countries on language-testing projects. Today, they are still at the forefront of some aspects of testing, working with rarely assessed languages for practical purposes and finding innovative ways to meet operational government needs.

Cross-References

- ▶ Criteria for Evaluating Language Quality
- ▶ Ethics, Professionalism, Rights, and Codes
- ▶ High-Stakes Tests as De Facto Language Education Policies
- ▶ History of Language Testing
- ▶ Testing Aptitude for Second Language Learning
- ▶ The Common European Framework of Reference (CEFR)

Related Articles in the Encyclopedia of Language and Education

Sally Magnan: The Role of the National Standards in Second/Foreign Language Education. In Volume: Second and Foreign Language Education

Chantelle Warner: Foreign Language Education in the Context of Institutional Globalization. In Volume: Second and Foreign Language Education

Wayne Wright, Thomas Ricento: Language Policy and Education in the USA. In Volume: Language Policy and Political Issues in Education

References

ASTM. (2011). F2889-11, Standard Practice for Assessing Language Proficiency, West Conshohocken: ASTM International. www.astm.org

Brau, M. (2013). ILR-based verbatim translation exams. In E. Galaczi & C. J. Weir (Eds.), *Exploring language frameworks: Proceeding s of the ALTE Kraków Conference, July 2011* (pp. 333–343). Cambridge: Cambridge University Press.

Brooks, R. L. (2013). Comparing native and non-native raters of US Federal Government speaking tests. Doctoral dissertation. Retrieved from WorldCat Dissertations and Theses. (Accession Order No. 867157336).

Brooks, R. L., & Mackey, B. (2009). When is a bad test better than no test at all? In L. Taylor & C. J. Weir (Eds.), *Language testing matters: Investigating the wider social and educational impact of assessment* (pp. 11–23). Cambridge: Cambridge University Press.

Brooks, R. L., & Brau, M. M. (2015, March). Testing the right skill: The misapplication of reading scores as a predictor of translation ability. Paper presented at the Language Testing Research Colloquium, Toronto.

Bureau for International Language Co-ordination (BILC). (2016). *Bureau for International Language Co-ordination*. Resource document. http://natobilc.org. Accessed 18 Aug 2016.

Child, J. R. (1987). Language proficiency levels and the typology of texts. In H. Byrnes & M. Canale (Eds.), *Defining and developing proficiency: Guidelines, implementations and concepts* (pp. 97–106). Lincolnwood, IL: National Textbook Company.

Child, J. R., Clifford, R., & Lowe, P., Jr. (1991). *Proficiency and performance testing*. Unpublished paper.

Herzog, M. (2003). An overview of the history of the ILR language proficiency skill level descriptions and scale. Resource document. http://govtilr.org/Skills/IRL%20Scale%20History.htm. Accessed 15 Apr 2015.

Higgs, T. V., & Clifford, R. (1982). The push toward communication. Resource document. http://eric.ed.gov/?id=ED210912. Accessed 2 July 2016.

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34*, 353–366. doi:10.1111/j.1745-3984.1997.tb00523.x.

Interagency Language Roundtable (ILR). (1985). *Interagency Language Roundtable Skill Level Descriptions: Speaking*. Resource document. http://govtilr.org/Skills/ILRscale2.htm. Accessed 15 Apr 2015.

Interagency Language Roundtable (ILR). (2016). *Interagency Language Roundtable*. Resource document. http://govtilr.org. Accessed 18 Aug 2016.

Jones, R. L., & Spolsky, B. (Eds.). (1975). *Testing language proficiency.* Washington, DC: Center for Applied Linguistics.

Kaulfers, W. V. (1944). Wartime development in modern-language achievement testing. *The Modern Language Journal, 28*(2), 136–150.

Lowe, P., Jr. (1985). The ILR proficiency scale as a synthesizing research principle: The view from the mountain. *Foreign Language Proficiency in the Classroom and Beyond, 17*, 9–53.

Lowe, P., Jr. (1988). The unassimilated history. In *Second language proficiency assessment: Current issues* (pp. 11–51).

Lowe, P., Jr. (1998). Keeping the optic constant: A framework of principles for writing and specifying the AEI definitions of language abilities. *Foreign Language Annals, 31*(3), 358–380.

Lowe, P., Jr. (2001). Evidence for the greater ease of use of the ILR language skill level descriptions for speaking. In J. A. Alatis & A.-H. Tan (Eds.), *Georgetown University Roundtable on Languages and Linguistics, 1999* (pp. 24–40). Washington, DC: Georgetown University Press.

Lunde, R. M., & Brau, M. M. (2005, July). *Correlation between reading and translation ability.* Paper presented at the World Congress of Applied Linguistics, Madison.

Lunde, R. M., & Brau, M. M. (2006, June). *Correlation between writing and translation ability.* Paper presented at the American Association of Applied Linguistics, Montreal.

Myron, H. B. (1944). Teaching French to the army. *The French Review, 17*(6), 345–352.

Office for the Director of National Intelligence (ODNI). (2016). *Foreign language*. Resource document. https://www.dni.gov/index.php/about/organization/foreign-language. Accessed 18 Aug 2016.

Petersen, C. R., & Al-Haik, A. R. (1976). The development of the Defense Language Aptitude Battery (DLAB). *Educational and Psychological Measurement, 36*, 369–380.

Pulliam, R., & Ich, V. T. (1968). The defense language proficiency tests: Background, present programs, and future plans. *Proceedings of the 10th Annual Conference of the Military Testing Association*, 69–82. Resource document. http://www.internationalmta.org/Documents/1968/Proceedings1968.pdf. Accessed 15 Apr 2015.