Cambridge University Press
978-0-521-16391-0 - Language Testing Matters: Investigating the Wider Social and Educational Impact of Assessment
Edited by Lynda Taylor and Cyril J. Weir
Excerpt
More information

1 When is a bad test better than no test at all?

Rachel L Brooks
Federal Bureau of Investigation
Beth Mackey
Department of Defense[1]

Abstract

Members of the US Government's Interagency Language Roundtable Testing Committee discuss Less Commonly Tested Languages (LCTLs) in the US Federal Government. The article focuses on the issues that arise when testing receptive skills and oral proficiency. The authors discuss how the US Government has adapted its traditional models of test development, validation, and tester training to provide solutions that satisfy both professional standards as well as its own institutional requirements.

Introduction

Gone are the days of the Cold War, when United States Government (USG) language testing organisations focused on a few, well-defined, visible languages, such as German and Russian, with sufficient, advanced resources. Today, intelligence sources come from every corner of the world and in hundreds of languages (NFLC 2005), many of which have never before been tested by the USG. With these new languages come increased challenges to language test development and administration, including difficulty locating appropriate resources, adapting testing instruments to the relevant culture, and standardising procedures across languages with efficient and effective methodology (Collins 2002) in order to accurately report on foreign language abilities.

Language testing is the gatekeeper for all foreign language intelligence and international interests, and the stakes are high for all parties involved (Laipson 2002). Decisions about which personnel are qualified to perform different language-related tasks depend on the validity and reliability of new tests being developed in languages that have never before been needed. Most directly, the quality of such tests affects the careers of the examinees: potential diplomats, agents, translators, interpreters, etc. USG language testers are charged with identifying these qualified language personnel, and the success of agency missions and the safety of non-language agency personnel depend on their language abilities. In turn, the security of the United States, and in some cases its allies, relies in part on the efficiency and effectiveness of USG agencies' testing practices.

When a USG testing department receives a new language testing request, testing specialists survey available resources and make decisions about how to best fulfil its requirements. Factors such as the timeframe of results, the number of people to be tested, the qualifications and availability of language experts to help develop and administer tests, the funding available for test development, and the applicability of the test to other agencies weigh in on test administration and scoring decisions. As the USG moves from testing in more commonly taught languages to less commonly taught languages, or even almost never taught languages, the challenges posed often mean that test development will be difficult, if not impossible, to undertake. Ultimately, USG testing organisations have no choice; we have to test. The question in the title, 'When is a bad test better than no test at all?' is not fair, because many times there is not a choice whether or not to give a test. The question really should be, 'How do we make the best test possible, given these conditions?'. This paper outlines the steps the USG is taking to produce reading, listening and speaking tests in a wide variety of less commonly taught languages, or perhaps more appropriately, less commonly tested languages (LCTLs).

Background

Many of the initial efforts to develop standardised language assessments in the USG have occurred under the auspices and direction of the Interagency Language Roundtable (ILR) (Herzog 2007). The ILR is an unfunded interagency organisation established for the coordination and sharing of information about language-related activities at the Federal level. It serves as the primary means of communication for departments and agencies of the USG, collaborating on issues of the progress and implementation of techniques and technology for language learning, use, testing, and other related topics. Despite its unfunded status, the ILR has made notable contributions to the language teaching and testing fields from its inception in the 1950s to the present (Chalhoub-Deville and Fulcher 2003, Clark and Clifford 1988, Herzog 2007, Lantolf and Frawley 1985, Lowe Jr 1988). Of particular interest to the language testing field are the Federal Government-wide Language Proficiency Skill Level Descriptions which detail an 11-level scale for foreign language skills of Speaking, Reading, Listening, and Writing (ILR 1985a, 1985b, 1985c, 1985d). Adapted from descriptions originally developed for the United States Department of State in the late 1950s, these ILR Descriptions have influenced the evaluation of foreign language proficiency both in the United States and internationally (Herzog 2007).

For about 50 years, the USG has used the language testing standards produced by the ILR to develop and conduct foreign language tests to meet the varying demands of its agencies. Initially, the demand was primarily for speaking tests in a fairly consistent set of languages including, but not limited to, Mandarin, French, German, Japanese, Korean, Persian-Farsi, Russian, and Spanish (Clifford and Fischer 1990). Well-educated, highly articulate native speaker testers were identified and methods for training testers were developed (Clark and Clifford 1988). Even in its seminal days, language testing across the Federal Government involved many languages that even today are not frequently taught or tested in academic and commercial contexts. Testing personnel in the USG have considerable experience with these languages and resources for testing in them are typically not terribly difficult to obtain, although recently there has been an increase in the number of tests administered (Tare 2006).

Today, the USG tests in well over a hundred languages for purposes as diverse as measuring the proficiency of diplomats in embassies, interpreters in courts, and soldiers on battlefields (United States 2001). Even though most of the languages sought by the USG today are not taught in the United States educational system, they are essential to operations and lives depend on personnel's ability to reliably use appropriate language skills. Daily decisions are made regarding how to best use the USG's limited resources, given agencies' differing requirements and priorities (NFLC 2005), and language test scores are regularly consulted to make those decisions.

USG agencies conduct tests in reading, listening, speaking, writing, translation, interpretation, listening summary translation, and other skills. Combined across language skills and agencies, the USG administers tens of thousands of language tests each year. Over 12,000 speaking tests alone are administered annually by the Defense Language Institute, the Federal Bureau of Investigation, and the Foreign Service Institute in over 100 languages. Over 500 testers in commonly and less commonly taught languages receive training, attend refresher workshops for re-norming, and undergo quality control checks continually.

In the past, USG test developers and tester trainers had been challenged to find resources in languages such as Hindi, Pashto, Persian-Dari, and Urdu, which still are not commonly taught in the US. Rarely tested by USG agencies in the past, these languages are now tested on a regular basis (Brecht and Walton 1998, United States 2001). Today, a new set of languages, including Baluchi, Sindhi, and Ibo (NFLC nd, United States 2001), impose new demands on USG testing personnel, necessitating adjustments to commonly used testing procedures to conduct language tests under conditions where time is short, resources are scarce, and accurate testing can be literally of life-and-death importance (The National Language Conference 2004). The development and administration of such new language tests by USG agencies means considering the nature of the language, the language's associated culture, qualifications of language subject matter experts, the language population, and test standardisation. A close examination of these issues, as well as the solutions the Federal Government has come up with to mitigate the problems, will lead to a more complete picture of the changing demands of language testing in the USG (Brecht and Walton 1998, The National Language Conference 2004).

Less commonly tested language issues

Language-specific issues

Anyone who has ever been required to make distinctions between languages has faced the difficult task of deciding what separates a language from related languages, dialects or other variations. Social and political circumstances affect language development, shift, and perception (Gordon 2005). These issues do not just trouble linguistic ethnographers, but also language testers. Test developers have important decisions to make about whether to test variations of a language separately, or as a single language. For example, in the past few years, the USG has moved from testing Serbo-Croatian as one language, to three distinct varieties: Serbian, Croatian, and Bosnian (Gordon 2005). Eastern Punjabi has been separated from Western Punjabi in USG testing. Each decision means redeveloping and validating existing language tests, dividing set resources, and re-evaluating previously established scores and procedures.

Many of the USG traditional testing formats are multi-level tests that include the top end of the ILR scale (Professional, Advanced Professional, and Native, or Levels 3 through 5). Some LCTLs may not be spoken at these higher ILR levels, or if they are, they might be combined with other languages. Some speakers convert to a different language altogether when raising the register, complexity, or sophistication of speech, often a colonial or standard language. Other languages may not simply switch to another language at a certain ILR level, but may switch languages in certain situations or during language tasks, causing the test language to be unable to meet all requirements of a particular level ILR description, as a proficiency test. Speakers of some languages often shift into another language when they move beyond the 'home and hearth' topics. For example, in the Philippines, speakers of Tausug or Chavacano shift either partially or completely into English and Spanish when topics increase in their level of abstraction (Gordon 2005). In some cases, the language does not change completely at the higher levels, but rather adopts a substantial amount of lexicon from another language.

For example Hindi incorporates English, and Cebuano incorporates both Spanish and English (Gordon 2005).

Another issue for consideration is whether every language can fulfil the description of ILR Level 5 without resorting to another language. Some agencies have determined certain language tests to cap off at Level 4, or in some cases, Level 3. Current discussion revolves around Arabic dialect testing, which switches to Modern Standard Arabic in certain contexts (Gordon 2005). Whether or not a single standard can be set for all Arabic dialects is debatable. As the USG increases its testing in the LCTLs, test developers have found the value in first describing the language testing requirement before tackling these issues, as an initial assessment of the purposes of the test results allows the USG to alter traditional formats to test only the pertinent ILR levels.

Determining when language interference is acceptable and when it is not can be difficult, as many languages may not have equivalents for foreign words. Sometimes the adoption of foreign words occurs only in specific subjects, such as technical fields. Language testers must consult with experts in the language to determine a standard for when and where foreign words are acceptable. Receptive skill test developers can avoid some of these pitfalls by omitting any potential test items that may require foreign words. Moreover, they can limit the range of levels that the assessment covers to the lower skill level descriptions, as long as the test fills the need of the relevant agency. Test raters for any open-ended items are trained on acceptable responses.

In productive skill assessment, Oral Proficiency Interview (OPI) testers are trained in how to handle language interference. Examinees are informed before the test starts that they should use foreign words only when they are a part of the language, and to avoid foreign words when target language words are available. If an examinee uses a word from another language or dialect, the tester should ask for explanation in the target language. Problems in communicating are resolved by both the testers and the examinee through circumlocution. Furthermore, periodic retraining is conducted and additional training provided before testing sessions to remind testers of issues of importance, good strategies to employ, and pitfalls to avoid. These sessions often include reminders of the language-specific strategies and language interference issues discussed in previously attended OPI training workshops. Testers are provided with written guides to use during the test, which reinforce the principles and procedures of speaking testing.

Cultural issues

When a USG testing organisation is tasked with developing, administering, and scoring a test in a language not tested before, the test developer not only has to be educated about the nature of the language, but also about the culture of the land where the language is spoken. Language and culture are inextricably entwined. One sociolinguistic and cultural issue in the target language is taboo topics. Topics that are acceptable for discussion in American culture may be considered offensive or personal in another, and vice versa. This issue needs to be managed very carefully by the tester. Even though test developers may be well-informed and prepared for any test in a new language, issues that present themselves during the course of the tests in LCTLs may be difficult to prepare for beforehand, as tester trainers may have limited knowledge of the language's social and cultural aspects. Experts in the language may know these details implicitly, but may not articulate them to test developers, until they come up in the course of the test.

Testers must also carefully consider the culture-specific appropriateness of particular speaking tasks and role-plays. It is important that the tester trainer be well informed of his or her participants' culture. A mistake like the choice of an inappropriate task or role-play may make the test seem biased or uncomfortable and potentially invalidate the results. Some languages' cultures have gender bias, where there are different expectations for the performance of a male versus a female. In other cases, the power imbalance between males and females may mean that the roles a female can play in a test are limited. As the party who gives the score, the tester has more power than the examinee in speaking tests. When the tester is a female and the examinee is a male, there is an imbalance of power in favour of the woman, which can make the testing experience uncomfortable or unacceptable in some cultures. Female testers from these cultures may find it difficult, if not inappropriate, to challenge men in the process of determining the linguistic ceiling during the course of a test, particularly if the result would be marked linguistic breakdown.

The same principle applies to issues of age and seniority. Younger testers may be hesitant to challenge older examinees, feeling that if they exposed the examinee's linguistic weaknesses, they would show lack of respect for the examinee. Likewise, the more senior examinee may be offended by the challenge from a younger member of the same society, or embarrassed by a weakness displayed during the test. Seniority in employment follows the same pattern. Instances occur when an examinee is a more tenured colleague of the tester. In some cultures, it would not be appropriate to challenge a colleague with more seniority in a way that might lead to embarrassment.

Individual qualification issues

The shift to LCTLs has also necessitated adjustments in the test development training. In the traditional language test development model, practitioners are highly trained, not only in the test language, but also in teaching and testing methodologies. As test development projects have shifted into new languages, locating qualified, educated speakers of these languages in the United States has proven difficult. Many of the target-language test developers have limited English skills; therefore, training such individuals in the ILR scale and testing models has required USG testing organisations to adapt their traditional models. The USG has found success in pairing highly experienced test development project managers who are adept at dealing with non-native speakers of English with native language consultants, who then work hand-in-hand to develop LCTL tests. This give and take between the language expert and the test developer is time consuming, but has produced successful tests in languages such as Dari and Pashto.

Testers used in LCTLs sometimes, if not often, do not have the ideal profile for language testing projects. They do, however, have the one quality that is irreplaceable, proficiency in the target language. The language testing organisations are challenged to find creative ways to overcome the lack of other necessary qualifications. Some of these tester recruits have no language teaching or testing experience, beyond what they experienced themselves learning a foreign language or undergoing language testing. Tester trainers are challenged to explain how tests are designed and function, and to undo any false perceptions about language testing, such as all native speakers always receive the highest score on the ILR scale. Tester recruits who perform at low levels in English in some or all language skills pose additional complications to trainers. It is sometimes difficult to discern if they internalised the USG standardised language testing system during the training, development, administration, and scoring. If so, to what extent did they grasp the necessary concepts?

Some of the individuals who are recruited have not lived in the country of the target language for decades. The constantly evolving nature of language leads to the possibility that the language as it is spoken in a country today has changed since the native speaker last lived there, and potential testers could be speaking an antiquated form of the language. Moreover, if the tester has been living in the United States for many years, it is possible that lack of practice in the language has caused attrition in the language. Attrition is particularly apparent in the sophisticated or complex speech of ILR Levels 4 and 5 because there are fewer opportunities to use a range of types of speech while living in the United States.

In situations where there is an urgent need and limited resources, USG testing organisations have discovered some options for assistance. Initially, other USG colleagues are consulted on information about the nature of the language, readily available language testing resources or current testing projects underway. Through the work of the ILR Testing Committee, members have developed partnerships across agencies that were not common in years past. A possible solution is to locate readily available materials or a trained, qualified tester at another agency. If none exist, efforts are made to locate employees within the agency who have the needed language ability on record. Such employees often do not have training in testing, but they usually have a security clearance, a certain amount of availability, and a willingness to help.

If there are no potential testers within the USG, some agencies are free to search for support outside the USG. Language communities in the United States often have community centres and organisations where resources can be found. Additionally, if the languages are taught at the university level, there are professors or other staff who can sometimes lend assistance. USG testing organisations have also tapped professional organisations, which often have a presence on the internet, linking speakers of a particular language who live in various parts of the country. Locating suitable testers is only the first step; the speaker must also be available and qualified to assist with testing.

In some cases, USG test developers and administrators have no choice but to use heritage speakers instead of native speakers. When heritage speakers are used to develop receptive test materials or administer speaking tests, the product may be flawed, threatening the outcome of the test. If the examinee has a higher level of speaking proficiency than the OPI tester, the validity of test results would be affected. Similarly, native speakers have varying levels of language ability, and a native speaker test developer or administrator may have a lower proficiency than the individual taking the test.

When new testers have little background in testing or underdeveloped language skills, individualised training has proven to be more beneficial than training in a workshop setting. Tester trainers can adjust the curriculum to the pace of the trainees, and meet their individual needs. Testers are closely monitored throughout training and test administration for new language-specific or testing theory issues that were not previously addressed, and additional training is provided to address any problems that may arise. If a trainee's native language is deficient in some way, available print and audio media can be used to update language skills. Active testers in all languages are also required to keep current in their language and practice to prevent attrition. Testers are encouraged to practise high-level language skills whenever possible.

In cases where trainees are deficient in English proficiency, an interpreter has been found to be of great use during the tester training. Chances are, though, that an interpreter would be difficult to locate, considering the initial difficulty in finding personnel with the needed language. As an alternative, an interpreter for a language closely related to the target language, perhaps spoken in the same area, can assist in training.

Careful and detailed explanation of what went on during the exam should be documented before assigning a score. Testers review all the tasks and topics posed to the examinee, including responses. Examiners pose detailed questions based on ILR levels, and help the tester to interpret the descriptions.

Only examiners who have an extensive testing and linguistics background should be used in these instances. Only examiners assign scores, though the tester may express an evaluation of what the score should be. Examiners take careful notes on the nature of the language and particular features of the language that were discovered through the course of the test. These notes are used for later reference in conducting tests by the agency and for training new testers in the future. Additionally, the notes can be shared with other agencies. Time and resources permitting, uncertainty in a test score can be resolved through a third party review.

To summarise, the fact that many of the native speakers have not had explicit training in linguistics or language acquisition is further complicated by the fact that all of this information must be relayed in English, which may be an area of weakness for the tester.

Population issues

In testing common languages, testing departments have followed traditional, large scale testing models that rely on piloting, validation and sophisticated item analysis of multiple-choice reading and listening tests. In the LCTLs, USG testing organisations often struggle to find an adequate number of people in the target population who can readily participate in a formal validation. Some of the LCTLs that need testing may come from populations of 10,000 to 100,000 speakers worldwide (Brecht and Walton 1998). The population in the United States to draw from for test development and validation projects is much smaller. In order to collect a large enough sample of speakers, the Defense Language Institute, for example, has had success in including both heritage and native speakers in the validation pool.

Diversifying the language validation pool creates its own set of challenges. Participants may have weak literacy and English comprehension skills; these deficiencies can result in the demands of the test not being met. Since finding large enough populations for thorough item analysis and calibration is difficult, constructed-response tests (CRTs) are being used when testing receptive skills. CRTs are somewhat more direct and flexible than multiple-choice tests, and protocols can be adjusted to accommodate novel examinee responses. The CRT format has been especially helpful in overcoming the difficulties of test developer qualifications and size of the validation population. CRTs are more time consuming to grade than multiple-choice tests, but the flexibility allows for a quicker development cycle. Possible test responses are collected, and using statistical analyses of the most likely and plausible responses, CRT items can be eventually converted to multiple-choice items.

Finding authentic materials in these new languages to be used for reading or listening tests can also be problematic. Media may largely be produced by a diaspora population not representative of the language as it is used

Cambridge University Press
978-0-521-16391-0 - Language Testing Matters: Investigating the Wider Social and Educational Impact of Assessment
Edited by Lynda Taylor and Cyril J. Weir
Excerpt
More information

in-country and, depending on the area in question, internet resources may not exist. Some government teams have found limited success in having test developers purpose-write passages for assessment purposes, but care should be taken that the language feels authentic. The USG has also tried to use a variety of diaspora sources, if diaspora sources are the only ones available.

Problems that plague receptive skill testing also affect the testing of speaking. In such testing, some OPI needs may be forecast well in advance, and others at the last minute. While USG agencies have developed capabilities in many languages over the past 15 years, in many other instances, there are no readily available resources, and speakers of that language may be difficult to find. Even when speakers can be found in the US, they may have spent so many years away from their native country that they are not in touch with the language as it is used today. Despite all of these difficulties, the OPI has become the *de facto* emergency language test, due to the fact that an OPI can be administered by trained native speaker via telephone to examinees in remote locations, and a previously prepared form does not have to be developed.

Standardisation issues

USG testing faces the need to create standardised proficiency tests across languages, with emphasis on the Middle Eastern, Central, and Southeastern Asian languages and their dialects. The more experience tester trainers have with testers of various languages and dialects, the better the understanding of how these languages function and interact with each other. Consequently, procedures for speaking testing across languages and agencies must be constantly re-evaluated. In particular, the ILR Skill Level Descriptions need to be applied to each USG test. We have discussed how the nature of these LCTLs is quite different from the languages government agencies are accustomed to testing. As new aspects of these languages emerge, we must interpret the Skill Level Descriptions consistently, to maintain test score reliability and validity. Procedures may evolve to meet USG agencies' changing needs; the language testing ethics and standards cannot be compromised.

Over the past several decades, foreign language test development across the USG has settled into traditional formats. For example, many reading and listening tests are linear, multiple-choice comprehension tests that measure a person's general ability to comprehend spoken or written language regardless of how it was learned, with reference to the ILR Skill Level Descriptions for Reading or Listening. Accordingly, OPIs conducted by USG agencies use the ILR Descriptions for Speaking. Studies of rating consistency have been repeated over the years. Moreover, the regular monthly meetings of the ILR present opportunities for individual agencies to share efforts to tackle pertinent language issues, as well as display advances made in language testing,