Assessment challenges in online instruction

Appropriate assessment in online learning

Martyn Clark, PhD, Christopher Green, PhD, Corey Miller, PhD, Karen Vatz, PhD, Medha Tare, PhD, Carrie Bonilla, PhD, Thomas Prado, MA, Evan Jones, MA

Assessment is a vital part of online learning. The use of assessment tools can provide stakeholders with information both to improve the effectiveness of instruction and to measure the impact of that instruction. The value of the information that a comprehensive assessment system can provide is directly related to the quality of thought that goes into the development and use of the assessment activities in that system. This is especially true for the LanguageNation platform, where assessment will be used not only for providing stakeholders with information about language training effectiveness, but also to provide the system itself with information needed to appropriately adapt the learning experience for individual learners and to improve its own functioning. Moreover, these assessment practices must be undertaken in the service of reaching language training program goals. And, although quality assessment instruments will produce more reliable information than poorly constructed ones, making better pedagogical decisions depends on collecting relevant information and not just reliable information. The goal is to produce learners who are proficient in the language, not merely proficient in the activities easiest to administer and assess online.

This report will cover the following topics that are especially relevant to LanguageNation:

- Essential assessment concepts
- Widely used proficiency scales
- Typical assessment uses in online courses
- A brief synopsis of Persian and Somali assessments available online
- An overview of promising avenues of exploration for incorporating assessment ideas into LanguageNation

This report is not intended to be exhaustive or definitive, as many aspects of the assessment activities eventually incorporated into LanguageNation will need to be developed in conjunction with the platform.

1 Portions of this report are adapted from:

Developing appropriate assessment tools is less a matter of “finding the right test” and more an issue of determining how to reliably capture the information needed for particular programmatic decisions, whether that decision involves determining the next activity for the learner or evaluating the effectiveness of an entire language learning program.

ESSENTIAL ASSESSMENT CONCEPTS

Before presenting some of the research relevant to assessment in general and online assessment in particular, it is useful to step back and consider the purpose of assessment activities. At the most fundamental level, assessment is the intentional collection of information for the purpose of making some kind of training decision (Norris, 2000). In other words, every assessment activity should produce information that helps to answer a question (e.g., “Should this learner be granted credit?” “Has the course sufficiently covered the material relating to this objective?” “Is this learner most likely to benefit from explicit instruction on particular grammatical points, or is a more integrated and implicit approach warranted?”). This is true whether the decisions are being made by a human curriculum coordinator or a computer algorithm. Depending on the type of decision to be made, the quality and scope of the information that is needed will change. Some decisions will only require a rough estimation of a learner’s abilities and, therefore, can be made with information that is somewhat imperfect or incomplete. In a simple branching algorithm to determine which of two possible instructional activities to present next, it may be sufficient to base that decision on the percent correct score from the previous activity. Other decisions, such as the evaluation of the overall success of the LanguageNation project, may have significant consequences attached, and, thus, will require information that is more extensive and precise.

A good language program relies on multiple assessments – multiple sources of information – to collect information that is appropriate for the particular decision that needs to be made (Brown & Hudson, 1998). The purpose for the assessment should drive the selection of assessment tools and procedures (Hughes, 2003). Multiple-choice or fill-in-the-blank items may be relatively easy to administer and score in an online environment, but may not provide sufficient information for every assessment need. For that reason, an imperfect measure of the right component may actually lead to better decisions and outcomes than a perfect measure of the wrong component.

Reliability, Validity, and Feasibility

Reliability and validity are fundamental concepts for describing the quality of an assessment instrument or, to be more precise, the quality of a particular use of an assessment instrument (Brown, 2008). Though a complete treatment of these concepts is beyond the scope of this report, a minimal understanding of these ideas is useful to recall in any discussion of assessment practices. Although feasibility is not an assessment concept per se, there are always real world constraints (e.g., resources, expertise) that might impact whether or not a particular assessment idea or tool can be implemented in a given context.

Reliability

Reliability refers to the consistency of the results of a given assessment instrument. Although there are various technical details involved in determining reliability of scores (see Thompson, 2003), the practical implications can be illustrated with an admittedly simplistic example often used when discussing the concept of reliability. Imagine the reliability of a bathroom scale and its effect on our decisions. If we were to weigh ourselves several times in a row on a “reliable” scale, we would expect the scale to display roughly the same weight each time. The scale would be of little use to us if it merely displayed a random weight each time we stepped on it. Tracking the progress of our diet would be an exercise in futility with an unreliable scale, as we would never know whether the changes we observed were due to actual changes in our weight or merely due to the unreliability of the measurement.

In the foreign language training context, we also want our assessment tools to be reliable. Tracking a learner’s progress using a particular assessment tool is only useful to the extent that the tool provides consistent scores given the same level of performance. The more reliable the tool, the more we can interpret differences in
the score as meaningful. This is true whether comparing the score of one learner to another learner at the same time, or comparing the current score of a learner with his or her score at a previous time.

For multiple-choice or other objectively scored tests, reliability is often expressed as a numerical coefficient, with values ranging from 0.0 (completely unreliable) to 1.0 (completely reliable). Some authors have suggested that typical reliabilities for an objective test of grammar or vocabulary should be .90 or above (McNamara, 2000). The reliability of high-stakes commercial tests will tend to be higher than the reliability of teacher constructed tests. For subjectively scored tests – such as essays, presentations, or proficiency interviews in which scores are awarded by raters – the percentage of agreement between raters (inter-rater reliability) or the consistency of a single rater over time (intra-rater reliability) may be calculated. (See Brown, 2005, for detailed descriptions of various reliability measures and instructions for determining reliability from a variety of tests using spreadsheet calculations.)

Two things should be noted here. First, there is always some amount of measurement error even with the most carefully constructed assessment instruments. For this reason, it is almost always better to use multiple sources of information (or multiple raters) for important decisions rather than rely on a single instrument (or a single rater). All other things being equal, longer tests tend to be more reliable than shorter tests because they provide more pieces of information (Allen & Yen, 2002). Second, reliability is not necessarily the same thing as sensitivity. To return to our weight example, just because a bathroom scale is reliable for weighing people to the nearest pound, that does not mean that it is sensitive enough to weigh letters for postage to the nearest ounce. In the same vein, just because a global proficiency test like the Defense Language Proficiency Test (DLPT) is reliable, the scores that it produces may not be granular enough to show small changes in ability over a short period of time. A proficiency test may not be appropriate for measuring incremental progress (Clifford, 2009). And, especially when measuring language development, treating all learner errors as “incorrect” when compared to a native speaker norm may fail to capture important aspects of progress.

Validity

Validity is generally concerned with the inference that one makes about the underlying ability of the people being tested based on their performance on a particular assessment instrument (Popham, 2009). In a driver’s license road test, for example, the ability to demonstrate various vehicular maneuvers (e.g., turning, parking, stopping) during the test is assumed to be representative of ability to perform similar maneuvers in other, non-tested locations, when driving different cars, on different days, etc. In other words, we want to generalize from what we can observe (the driver’s specific performance on the road test) to things that we cannot observe (future performance in driving based on an underlying “driving competence”). Thus, validity is determined, in part, by the quality and appropriateness of the content of the test. If the bulk of our assessment activities are focused on content that is unrelated or only weakly related to our ultimate criterion of interest, it will be difficult to predict future performance accurately no matter how reliable the assessment itself.

To put this in a foreign language training context, a proficiency test is only useful to the extent that the score from such a test provides a reasonable estimation of the test-taker’s real world functional ability to use and understand the foreign language when confronted with actual foreign language use situations. We make an inference from the test performance to the real world ability. Those people who scored higher on the test should be able to demonstrate greater functional ability in the “real world” than those people who scored lower on the test. To the extent that those inferences hold, we think of the test as a valid indicator of proficiency.

The more direct an assessment (the more it faithfully elicits the target activity of interest), the clearer the link between the assessment activity and the intended inference from the score. However, because it is not always easy to directly measure those things that we are most interested in, there can be a tension between reliability and validity as we move along a continuum from more direct to less direct assessment. It may be possible to score a multiple-choice vocabulary test more reliably than a test that requires the student to give a

---

2 Kane (2010) argues that not only is measurement error present, it is desirable if one intends to generalize beyond the particular conditions of the observed performance. We tend to assume that there is some generalizable construct (e.g., language proficiency) that underlies performance across tasks, and that minor performance inconsistencies for the same learner in the very short term (e.g., minutes) do not represent substantial changes in the underlying construct.
mock briefing, for example, but the mock briefing may be a more meaningful indicator of functional language ability and more in line with the ultimate goals of a language training course. At the same time, however, if the instructor or algorithm is wildly inconsistent in scoring the mock briefing across students, any inferences we try to make from those scores will be suspect, as we cannot be sure if score differences are due to the actual underlying ability or merely to the unreliability of the scoring process.

Unlike reliability, there is generally not a single number that is used to quantify the validity of a test. Instead, a careful description of the test’s purpose, content development, and appropriate score uses are woven into a narrative known as a validity argument (see Kane, 1992) or a test use argument (see Bachman, 2005). These arguments lay out the conceptual link between the assessment tasks (e.g., the items on the test) and their relation to the underlying abilities that they are intended to measure (e.g., listening proficiency, vocabulary knowledge). Ideally, in addition to clearly detailing the conceptual links between the assessment tasks and the appropriate score inferences, empirical evidence is collected to support the validity argument.

Test validity arguments are based on a series of logical connections between the items or tasks on the test and the meaning of the test scores. As such, test validity arguments are always defeasible; that is, there is always the possibility that additional information will weaken the argument. In recent years, there has been much use of Toulmin’s (1958) argument structure as the basis for developing validity arguments (e.g. Mislevy, 2007). The basic structure of the argument is shown in Figure 1. Data (e.g., performance on a given assessment instrument) is used to make a claim about an examinee (e.g., that the examinee has a particular language proficiency level). This requires an inference from the data to the claim, which is supported by the warrant (the logical statements that serve as a bridge between the the claim and the data; e.g., the test in question is a reasonable measure of proficiency). The warrant, in turn, is supported by backing (statements which give reasons to believe that the warrant is true; e.g., a discussion of the elements of proficiency and how those are embodied in the test design). The overall claim is subject to rebuttals, which may place some qualifiers that limit the strength of the argument or give specific conditions under which the argument holds (e.g., assuming that the test was given under appropriately proctored conditions with trained raters). An example of this approach applied to the Test of English as a Foreign Language (TOEFL) can be found in Chapelle, Enright and Jamieson (2010). Another example applied to the High-level Language Aptitude Battery can be found in Mislevy et al. (2010).

For instructor-developed assessments or assessments based on online instructional activities, the logical connections between assessment tasks and the meaning of a particular learner’s performance on those tasks may be implicitly (and unproblematically) assumed rather than explicitly developed in a formal validity argument. Nevertheless, the same underlying logic applies.

Feasibility

Apart from the inherent tension between validity and reliability, there is a layer of practical feasibility in the implementation of any assessment. A computer-mediated oral interview, for example, requires reasonable schedule availability of the participants in addition to having a well-documented interview protocol and a well-trained interviewer. In many language programs, technical, practical, or financial considerations may play a substantial role in the final selection of assessment activities. However, although feasibility will be a mediating factor in the final decision, it is of the greatest importance to try to clarify the goal of the assessment before letting practical constraints push the discussion towards a particular solution. For example, just because the

---

3 There does exist a formula for a validity coefficient, based on a test’s reliability, but this is rarely encountered in most discussions of validity. That validity coefficient does suggest, however, that a test’s reliability sets the upper bound on its validity (Hening, 1987).
4 “[E]ach testing context in any given time and place has functional, political, or economical constraints that need to be considered in terms of what can actually be accomplished with regard to changing tests and testing policies, curriculum and curriculum policies, educational priorities and policies, and so on” (Brown, 2008, p. 295)
current limitations of automatic speech recognition may preclude the development of an on-demand, computer-scored interview test, it does not follow that an easy to implement multiple-choice test is an acceptable alternative. Circumstances may dictate the use of a less than optimal solution, but this should be a reasoned decision; expediency should never be the primary criterion.

Norm-referenced and criterion-referenced decisions

Norm-referenced decisions and testing (NRT)

In addition to the core notions of reliability and validity that underpin all test score interpretations, there are some additional concepts that provide ways of thinking about the relationship between test scores and educational decisions. Language tests can be classified into two basic families depending on whether the decision to be made is norm-referenced or criterion-referenced (Brown, 2005). A norm-referenced decision is one in which the performance of a given test-taker is measured against that of other test-takers. The goal for a norm-referenced test, generally speaking, is not to determine the absolute ability of any particular test-taker, but rather to spread test-takers out along a particular continuum of ability in order to identify those who have relatively more or relatively less of that ability. The SAT,\(^5\) which measures general academic readiness for college, is a well-known example of a norm-referenced test. The score on the SAT represents how a particular test-taker’s performance compares to the performance of the population of test-takers. An SAT score in the 98th percentile would mean that the test-taker did as well or better than 98% of the students who took the test. Thus, the score is meaningful by virtue of its relation to other scores on the same test.

For a language program, placement testing is often at least partially a norm-referenced decision, especially when there are many learners to be placed into classes. Because classes that have learners with relatively

---

\(^5\) SAT used to stand for the Scholastic Aptitude Test. However, the test is now known as the SAT Reasoning test. Many aptitude tests are norm-referenced, as the goal is to spread examinees out on a continuum, from relatively less aptitude to relative more aptitude in a particular domain.
homogeneous abilities are usually easier to teach than those with learners of disparate abilities, knowing which learners “clump” together along the ability continuum might be useful information. That is, even within the same larger ability band, a program may choose to have one class section composed of “stronger” learners and another of “weaker” learners or, conversely, may want to ensure the full range of abilities in each class and intentionally pair stronger and weaker students during activities.

As the number of users in LanguageNation increases, norm-referenced measures of language ability could potentially be exploited to create virtual cohorts of learners of various ability combinations for virtual pairwork or peer-feedback. Norm-referenced measures can also be useful in exploring variables potentially relevant to platform usage (e.g., frequency of access, length of access, number of tasks attempted in single session). Identifying and interviewing users at the extremes on those variables (i.e., far above average, far below average) could provide insight into factors that affect platform usage. Note that during the initial stage of implementation, the norm-referenced population will need to come from outside of LanguageNation, but over time, the accumulation of information across LanguageNation users will ultimately allow the system to self-generate norm-referenced performance metrics from the user base.

Criterion-referenced decisions and testing (CRT)

A criterion-referenced decision, on the other hand, is concerned with the degree to which a test-taker has achieved a particular level of performance on specific, well-defined criteria (Popham, 2003). The goal of a criterion-referenced test, therefore, is to provide information about the specific capabilities of the test-taker. In general, the criteria on which the examinee will be judged are known to him or her in advance. The knowledge portion of a driver’s license test, for example, is composed of questions designed to assess the test-taker’s knowledge of the various rules of the road. Obtaining a “passing” score on the test depends on how well the test-taker can demonstrate his or her knowledge (e.g., at least 80% correct), irrespective of how well or how poorly other test-takers have performed. For a criterion-referenced test, the score is meaningful by virtue of its relation to the mastery of a particular skill or body of knowledge.

In a language training context, a criterion-referenced approach to assessment would mean that the goals of a particular instructional sequence would provide the content for the assessments. This suggests that assessment development and curriculum development go hand in hand, as it is not possible to assess performance on specific, well-defined criteria in the absence of determining what those criteria should be. This iterative approach to curriculum and assessment design is illustrated in Brown (2005) with the example of criterion-referenced testing in a university level ESL program.

Although both norm-referenced and criterion-referenced tests have their place, adopting a criterion-referenced point of view is beneficial for language training programs, as criterion-referenced testing was designed to measure instructional change (Glaser, 1963). If the goal of a particular course is to have the learners master certain skills or develop the ability to perform tasks or other instructional objectives, it makes sense to measure progress according to those skills, tasks and objectives. The “backwards design” approach to curriculum development operates in this fashion (Wiggens & McTighe, 1998).

Criterion-referenced testing is especially useful for pre/post measurement of gain. Imagine the following idealized scenario. Prior to the start of a course, a test is constructed to assess performance on the final learning objectives of that particular course. If the students take the test on the first day of class, one would expect them to perform rather poorly, as they have not yet mastered those particular objectives and, thus, have need for instruction. On the other hand, if the same test were given at the end of the course, one would expect the students to perform very well if the instruction had been successful.6

---

6 Because of the lopsided nature of CRT scores, there are some conceptual and statistical differences in determining the reliability of criterion-referenced tests compared to norm-referenced ones -- see Brown and Hudson (2002) for discussion.
Summative versus Formative assessment

Broadly speaking, a given assessment tool can be classified as providing primarily summative or primarily formative information. The goal of a summative assessment is to provide an evaluation of a learner’s knowledge or ability at a particular point in time, usually for the purposes of accountability. The DLPT, for example, would be primarily considered a summative assessment. The goal of the DLPT is to identify where the test taker falls along the proficiency continuum according to the Interagency Language Roundtable (ILR) scale in the particular skill being tested. (A later section describes this scale in more detail.) In addition to being a summative assessment, the DLPT can also be considered a high-stakes assessment, as the results of the assessment can have a great impact on the test-taker, such as the awarding or withholding of FLPP (foreign language proficiency pay). Summative assessments do not necessarily need to be external to the instructional context, however. Final examinations or unit quizzes in classroom instruction can also be considered assessment for primarily summative purposes, as their goal is to determine how well a learner has mastered the material from a particular part of the class. In the case of a final exam, the test would be (ideally) both summative and criterion-referenced.

In contrast to summative assessment, a formative assessment is one in which the results of the assessment are intended to provide feedback to directly improve the learning process. As such, formative assessments do not necessarily produce grades or scores in the traditional sense, as the provision of feedback to both the learner and the instructor is the primary focus. It is important to note that language learners do not progress immediately to mastery upon instruction. This means that adequate learner performance may be defined quite differently at different points during instruction, and acceptable (expected) performance from a formative point of view may initially be quite different from native speaker norms. Developmentally sensitive measures of progress must be chosen so that learner improvement can be accurately monitored. (See Norris & Ortega, 2003, for a detailed discussion of measurement issues in SLA.)

In recent years, there have been increased calls to extract formative information from assessments traditionally considered summative (see Huff & Goodman, 2007). This is part of the growing recognition that both classroom assessment, as well as more formal assessment, work best in tandem to provide a more comprehensive picture of student learning (Pellegrino, Chudowsky, & Glaser, 2001). For LanguageNation, adaptivity at the individual learner level might be primarily driven by formative measures sensitive to language development and learner needs; summative assessments of language proficiency (and various platform usage metrics) would provide information relevant to the overall evaluation of the project.

LANGUAGE PROFICIENCY SCALES

Before looking more closely at assessment tools in practice in foreign language courses, it is useful to briefly describe how language proficiency is defined in the U.S. Government and K-16 contexts.

The development of global language proficiency assessment in the U.S. is relatively recent, and originated in the U.S. Government. The Foreign Service Institute/Interagency Language Roundtable (FSI/ILR) scale was the first language proficiency framework designed to provide an objective rating system of language ability for U.S. Government employees. The FSI/ILR was originally developed by FSI in the 1950s for the purpose of creating an inventory of language proficiency among government employees. Over the years, the scale has been revised, with major reiterations focusing on standardizing rating factors, thus reducing subjectivity and improving reliability. Today, the ILR scale is used widely throughout the USG, is applicable to all languages, and has served as the basis for academic proficiency guidelines developed by the American Council on the Teaching of Foreign Languages (ACTFL; Clark, 2013).

The ILR scale comprises six major proficiency levels, ranging from 0 (No Proficiency) to 5 (Functionally Native Proficiency), with “plus-levels” assigned when an individual’s proficiency exceeds the skill criteria of one level, but does not quite meet the criteria of the subsequent level. Each level, including the “plus-levels,” is

---

7 For additional information concerning the ILR scale, please reference http://www.govtilr.org/skills/index.htm.
defined by a description of the typical language ability profile of a person at that level; separate descriptions have been adapted for the primary skills of reading, writing, speaking, and listening, and more recently, for the skills of translation, interpretation, and intercultural communication. The reading and listening descriptors are accompanied by a passage rating system that was developed to assess the complexity (e.g., concrete, factual language vs. nuanced opinion) of a text or passage intended for native speaker use (Child, 1998).  

In the 1980s, the Educational Testing Service (ETS) and ACTFL developed a set of proficiency guidelines based on the ILR scale definitions (Hiple, 1987). The ACTFL guidelines are intended for academic use, and like the ILR scale, have been refined over the last two decades. Currently, the ACTFL proficiency guidelines are divided into a five point scale (novice, intermediate, advanced, superior, and distinguished), with sublevels of low, mid, and high further defining novice, intermediate and advanced. In order to make this scale more usable for the intended population of typical foreign language students in the U.S., in contrast to U.S. Government professionals, the descriptions at the lower end of the scale were expanded to create a more finely graded scale that would allow for greater differentiation among the lower proficiency levels. Furthermore, the numeric system was replaced with user-friendly labels (Liskin-Gasparro, 2003).

The oral proficiency interview (OPI) is a widely used, standardized assessment tool designed to elicit language samples that can be matched to descriptions of a language proficiency scale, typically the ILR or ACTFL scale. The interview is conducted in the foreign language by a rater trained to probe the interviewee’s foreign language ability through a structured, yet unscripted, interview that unfolds like a conversation (Clark & Clifford, 1988). The speech samples obtained are then assessed according to the scale level descriptors for each of the scale levels, with the goal of choosing the scale level which best matches the sample. If the scale is non-compensatory – such as the ILR scale – the proficiency rating is the highest level at which a candidate can demonstrate sustained performance.

The ILR proficiency level descriptors and ACTFL proficiency guidelines are widely used throughout the U.S., however, other proficiency scales also exist. The Common European Framework of Reference (CEFR) has been equally influential in developing a framework that defines language proficiency (Council of Europe, 2001). As with the ILR and ACTFL approach, the CEFR levels are accompanied by descriptions of proficiency that range from basic (A1) to mastery (C2). The CEFR is a framework and not just a proficiency scale and, as such, includes a wide range of ancillary materials intended to assist with curriculum development and instruction.

Common to all of these approaches is the treatment of language proficiency as a functional rather than knowledge-based phenomenon. Proficiency is the ability to use language knowledge for real-world purposes. For this reason, discrete tests of language knowledge, such as vocabulary tests or grammar tests, are not sufficient for assessing proficiency, though they may correlate with proficiency measures. In many cases, the job of assessing oral proficiency falls to human raters who have been trained in both the descriptors which delineate the various proficiency levels as well as the appropriate elicitation techniques for probing specific aspects of the examinee’s functional ability at each level. The inherently adaptive nature of proficiency testing lends itself to a computer-delivered format; however, until human language technology (HLT) techniques are sophisticated enough to mimic the thought process of trained raters, it is impossible to fully take advantage of this natural synergy.

ALIGNING ASSESSMENT AND INSTRUCTION IN ONLINE PLATFORMS

The assessment of student learning is an integral part of second language (L2) training, including online learning. The decisions needed in online foreign language programs are not so different from those in a traditional face-to-face programs. The following section briefly presents the types of decisions that may be needed in assessing

---

8 Both ILR- and ACTFL-aligned tests use Child’s (1998) passage rating scheme.
10 The most recent version of the ACTFL Guidelines, including video samples of performance at various levels, is available at http://actflproficiencyguidelines2012.org/.
online learning (based on Hughes, 2003). They are presented chronologically, with those appearing earlier in the list tending to occur earlier in the overall learning sequence. A sampling of relevant research is included in this discussion, some of which uses commercially available assessment instruments. To support the Year 1 and Year 2 LanguageNation development work in Persian Farsi and Somali, a detailed analysis of the assessments currently available in those languages is included in Appendix C.

As Blake (2009) indicates, with only a few exceptions, the existing empirical studies on online learning do not measure L2 outcomes, possibly due to logistical issues associated with assessment at a distance. While it is a simple matter to assess learner performance on basic grammar rules or reading comprehension questions with self-correcting, computer-based exercises, it is much more difficult to determine how well learners perform more complex tasks in the target language, such as participating in novel conversations with fluent interlocutors or summarizing texts and audio recordings via use of audio- or video-based instructional technology; in fact, such assessment often exceeds the technological capabilities of automated tools. Therefore, there is a need to consider effective methods of assessing learners participating in online learning foreign language courses within a comprehensive assessment framework designed to systematically document their progress in online courses and evaluate the effectiveness of the courses over time. Because the LanguageNation vision is one of tailoring language instruction to the individual learner, the ongoing assessment of learner progress is a vital component in enabling system adaptivity.

Decision types in online instruction

**Screening**

Not necessarily a test of language ability *per se*, a screening decision is one in which the goal is to determine whether someone has the appropriate prerequisite knowledge or background to continue with a certain task, in this case with taking an online course. To make the decision, it may be desirable to collect information about the learner’s access to a computer with internet connection, familiarity with computer technology use, and ability to work with minimal supervision (Winke & Goertier, 2008). There may also be scheduling availability considerations if different time zones are involved.

**Diagnostic**

The purpose of various diagnostic tests is to get information to help match the instruction to the particular learner’s ability profile. Relevant information could potentially include a detailed assessment of the learner’s aptitude for learning foreign languages in general. For example, the High-Level Language Aptitude Battery (Hi-LAB; Doughty et al. 2007) measures an individual’s cognitive and perceptual aptitude and is built upon the following constructs hypothesized to be relevant to adult language learning: executive functioning (working memory, inhibitory control, task-switching), rote memory, perceptual acuity, processing speed, primability, and implicit and explicit induction. See Tare et al. (2013) for a discussion of individual characteristics such as these which can be used to build a learner profile. Similarly, information pertaining to a learner’s current foreign language skills can be used to tailor instruction to fit the particular individual.

Whereas Hi-LAB measures tap the cognitive underpinnings of adult language learning, other diagnostic measures consider current language ability and future language needs. The Online Diagnostic Assessment (ODA; see below) or the diagnostic interview (Cohen, 2003) are examples of this kind of information. Pre-course needs analysis (Long, 2005) would also fall under this general category. Initially determining learners’ various proficiency levels through Diagnostic Assessment (DA) is an important first step when designing an online learning course. DA tools allow instructors to properly tailor course content to specific students’ needs, as

---

11 The opening chapters of Alderson (2005) provides a fairly pointed critique of most current foreign language diagnostic testing. His main contention is that there has been insufficient differentiation between diagnostic tests and other, more general, language tests. Since part of the goal of diagnostic testing is to determine *specific* remedies or courses of action based on the results, the tests must be based on some conceptualization of the types of problems that a given learner might have and how those particular problems might be remedied, given a particular underlying theory of language development.
well as identify additional assessment tools that will appropriately document learner growth and progress during the course. Institutions often use locally developed DAs, tailored to the unique instructional goals of their language learning program, to obtain initial estimates of learners’ language abilities (two are discussed below). However, a number of web-based DAs are also available and can potentially be readily incorporated into an online learning assessment framework.

The Defense Language Institute Foreign Language Center (DLIFLC) offers ODAs in Arabic, Chinese, Dari, Farsi, Korean/North Korean, Russian, Spanish, and Tagalog. Reading and listening ODAs are available for Arabic, Chinese, Korean, and Russian, while the ODAs for Dari, Farsi, Portuguese, Spanish, Tagalog, and are available for reading only. The ODAs are available free of charge and posited to help individuals manage their language learning through the provision of individualized feedback specific to the reading and listening skill domains. The reading sections of the ODAs, encompassing authentic material in the target language and consistent with ILR-level descriptions, require learners to provide answers to comprehension and vocabulary-related questions concerning grammatical, discourse, and lexical passage characteristics. The ODA listening sections require the learner to listen to either authentic or modified versions of a given passage in the target language (up to two times) and answer comprehension questions in a free response format. For both the reading and listening sections, questions are presented to the learners in English, and learners are required to answer in English, as well. The DLIFLC ODAs, thus, are designed to reflect the type of tasks and response format a learner can expect to encounter when taking a DLPT. Although questions are presented in an open-ended format, they are designed to elicit a narrow array of correct responses. Keyed responses are then compared by the computer against a set of pre-identified “protocols” of acceptable correct responses. The DLIFLC ODAs are semi-adaptive; advanced learners are provided with more passages and questions than novice-level learners. ODAs are estimated to take between one and two hours to complete, depending on a learner’s level; upon completion, feedback is provided to learners in the form of a diagnostic profile, listing the tasks (such as reading comprehension or vocabulary-related items) successfully completed as well as information concerning areas in need of improvement. The lack of information concerning how raw ODA scores relate to known proficiency scales, such as the ACTFL or ILR proficiency scales, potentially limits the generalizability of ODA outcomes within a broader online learning context. That is, instructors who intend to use ODAs for a large number of learners as a pre-instructional tool for gauging students’ initial proficiency levels will be required to analyze each learners’ diagnostic profile and make inferences concerning learners’ starting proficiency levels. Since the material is designed relative to ILR-level descriptions, instructors must infer that successful task completion relates to the ILR level associated with the task(s) (see http://oda.lingnet.org/ for a description of how assessment texts are selected). However, detailed information concerning how a learner’s performance on an ODA relates more broadly to his or her overall proficiency level is not readily available. For instructors working with just a few learners, or for learners simply interested in obtaining a diagnostic foreign language profile for their own edification, this may not be problematic.

CASL researchers are currently working on developing sensitive measures to define the elements of interlanguage development at the different levels of language proficiency on the ILR scale. The goal of this project, Linguistic Correlates of Proficiency (LCP), is to gather data on different aspects, for instance, linguistic features, of the interlanguage of speakers at varying proficiency levels, in order to identify which features learners of varying levels master and which features remain difficult. A proof of concept study of a battery of Russian perception and production tasks found that various tasks measuring control of different aspects of Russian phonology, morphology, syntax, lexis, and collocations do differentiate between ILR levels 2-3 (Long, Gor, & Jackson, 2012). Establishing which linguistic features learners need to acquire at varying levels of proficiency can help inform and tailor instruction and underpin adaptivity. Test batteries are currently in the piloting phase for three other languages: Farsi, Mandarin Chinese, and Spanish. The tests are currently being delivered in a remotely executable format using DMDX (Forster & Forster, 2003), an experiment presentation software program.

12 For additional information, see http://oda.lingnet.org/.
13 Interlanguage is the dynamic but systematic continuum between the learner’s first language (L1) and second language (L2) that contains structures from both the learner’s L1 and L2 (Selinker, 1972).
Placement

When there is a wide variety of courses available, the goal of a placement decision is to match the learner with the most appropriate course. In a program with adaptive course content, such as LanguageNation, this could also involve determining the appropriate starting point. Alternatively, when there is a large number of learners taking similar courses or when learners will need to participate in online activities, the placement decision may focus on identifying groups of students with similar profiles or levels. In larger groups, it can be difficult for the instructor to tailor the instruction if there is wide variation of ability among learners. For effective placement, information about both the learner and the available course options are needed to make this decision. If there is a mismatch between the needs of the learners and the contents of the courses, no placement test will be able to provide an adequate match.

Since the instructional goals of online language courses vary widely in focus—from academically rigorous courses intended for professional certifications or college credit to those providing opportunities for motivated and interested learners to maintain their language skills—instructors often are interested in determining the estimated pre-course proficiency levels of their learners, particularly when previous outcomes data are not available or no longer current. One way to match course content to specific students’ needs is to administer a curriculum-driven placement test. For example, Brigham Young University (BYU) uses Web-based Computerized Adaptive Placement Exams (WebCAPE) for Chinese, French, German, Italian, Spanish, and Russian to estimate learners’ ability levels at the outset of each semester, although it should be noted that the assessments themselves can be given at any time. Since the tests are computer adaptive, the number of questions each learner answers is dependent upon proficiency level. The raw scores generated for each learner are then measured against cut-off scores to determine the placement of each student into an appropriate course level. Bulk site licenses are available for purchase via BYU’s Human Language Technology and Research Support Center website.14 As with the DLIFLC’s ODAs discussed above, no information is provided concerning the relationship between WebCAPE outcomes and established proficiency scales, limiting the extent to which meaningful inferences can be made between placement test outcomes and their pre-course general proficiency levels, which may be of interest to instructors working with learners at a variety of ability levels. While WebCAPE may be a helpful course placement tool for a variety of foreign language learners and programs, it may not necessarily benefit language programs that do not offer a complete suite of sequenced online learning courses (i.e., novice through intermediate). Finally, WebCAPE provides no meaningful feedback to learners concerning specific areas of test performance or skill-areas in need of improvement.

The DIALANG system, administered by Lancaster University, offers web-delivered ODAs for reading, writing, listening, grammar, and vocabulary in 14 European languages (Danish, Dutch, English, Finnish, French, German, Greek, Icelandic, Irish Gaelic, Italian, Norwegian, Portuguese, Spanish, and Swedish).15 DIALANG is currently free of charge for all languages and skill modalities. The assessments were developed to provide learners with estimated proficiency levels consistent with the Common European Framework of Reference (CEFR), and results are, therefore, reported relative to the six levels of the CEFR scale.16 Before beginning the actual DA, learners have the option of taking a Vocabulary Skills Placement Test (VSPT) and answering self-assessment questions. For learners who choose to take either the VSPT or self-assessment, the DIALANG system will use either or both of the test results to determine the appropriate level of the diagnostic test to administer. Otherwise, a medium-difficulty version of the selected test is administered as the default. Each DIALANG test session is expected to last between 30–45 minutes, including the time necessary to complete the VSPT and self-assessment tests. In contrast to DLIFLC ODAs and WebCAPE, learners are provided with the option of receiving immediate feedback after the completion of each test question, an option which can be enabled or disabled at any time. Upon test completion, learners are provided with a CEFR estimate of their language proficiency level, feedback on their performance within each sub-skill, and information concerning the extent to which their self-assessment estimates corresponded with their performance on the DA, in addition to descriptions of the ability levels immediately above and below their current estimated level of proficiency.

14 See http://creativeworks.byu.edu/htsrc/LicenseAgreement.aspx.
15 See http://www.lancs.ac.uk/researchenterprise/dialang/about).
16 For additional information concerning the CEFR scale, please reference http://www.coe.int/t/dg4/linguistic/cadre_en.asp.
Current limitations associated with use of the DIALANG system, however, are that it is limited to European languages, and, although some initial work has reported outcomes on pilot testing and validity research (Alderson & Huhta, 2005), DIALANG outcomes relative to the CEFR scale are not yet validated for all languages and skill modalities.

**Achievement**

An achievement or progress decision is intended to assess to what degree the learner has mastered the material that has been taught. Ideally, this is a decision made with reference to course goals in a criterion-referenced fashion. This can be at the level of the entire course and be primarily summative, as in deciding whether or not to grant course credit. Alternatively, the this can be more narrowly focused and used primarily formatively to determine how the instruction should proceed. For training programs in which the ultimate course goal is to achieve a particular level of proficiency as determined by a standardized proficiency test, the achievement test may, in fact, be a proficiency test.

**Proficiency**

The ultimate goal of most foreign language instruction is to produce learners who can actually use the language to perform real world tasks, such as having a conversation with a native speaker of the language, giving a briefing in the language, understanding a news presentation in the language, and the like. The term “performance” is sometimes used rather than “proficiency” when referring to those functional situations in which specific training has been undertaken (Clifford, 2009).

A comprehensive search for commercially available online assessment tools capable of documenting learner progress identified eight potentially relevant instruments, including the following: (1) Computerized Assessment of Proficiency (CAP), (2) the Standards-based Measurement of Proficiency (STAMP), (3) the ACTFL Assessment of Performance toward Proficiency in Languages (AAPPL), (4) Pearson’s Versant Oral Proficiency Test (OPI), (5) the Center for Applied Linguistics Computerized Oral Proficiency Interview (COPI), (6) ALTA Language Services proficiency tests, (7) Minnesota Language Proficiency Assessments (MLPA), and (8) Defense Language Proficiency Test Very Low Range tests (DLPT VLR). Appendix A provides an overview of the key components associated with each of the assessment instruments. (See Appendix C for assessments immediately relevant to LanguageNation.)

As can be gleaned from Appendix A, each of the computer-based proficiency tools currently available has both advantages and limitations associated with potential outcome measures for documenting learner proficiency gains during and after participation in an online learning course. Overall, few of these instruments are granular enough to capture learner proficiency gains after participation in short-term courses. In addition, one of the main controversies in the broader L2 testing literature—concerning quantitative computer-based testing (CBT) or computer-adaptive testing (CAT)—involves the effectiveness of computers, humans, or both in scoring L2 output in various skill modalities (Attali & Burnstein, 2006; Chun, 2006; Lafford, 2009; Ockey, 2009). Applied specifically to the online learning context, the CAP test implementation would, to take one example, require a substantial amount of effort to train raters in reliable use of the rating scales to assess learner output for both speaking and writing. While the Versant test does not require rater training, incorporation of its use within an online learning assessment framework is limited to the assessment of speaking skills only, although the developers claim that the results correlate and predict reading skills as well. The Versant’s use of

17 Although the Pearson website describes the Versant assessment instruments as proficiency tests, few empirical studies to date have examined the extent to which outcomes on the Versant test are predictive of outcomes on higher-stakes proficiency tests, such as the DLPT or OPI. For example, empirical research by Berman et al. (2008), designed to investigate the extent to which Versant test scores reliably predicted end-of-program outcomes, showed the Versant test to be a better predictor of speaking skills than reading or listening skills, as measured by the OPI and DLPT, respectively.

18 A comprehensive database of paper-and-pencil, computer-delivered, and computer-adaptive assessment instruments is maintained by the Center for Applied Linguistics (CAL) and can be accessed through the following link: [http://www.cal.org/CALWebDB/FLAD/](http://www.cal.org/CALWebDB/FLAD/).
ASR to score learner speech samples automatically necessarily constrains the types of test items included within the assessment instrument to decontextualized language activities, such as vocabulary repetition and sentence reorganization tasks. The current emphasis on providing learners with opportunities to engage in authentic communicative exchanges argues against implementing an assessment instrument that lacks face validity regarding genuine online foreign language interaction. Nevertheless, the short time required to complete the assessment, as well as the ease with which the test may be administered, may make the Versant instrument an attractive assessment tool for inclusion within an online learning course.

Although a number of studies have reported findings concerning the use of online proficiency-based assessment instruments in a variety of instructional contexts (Berman et al., 2008; Burwell et al., 2009), to our knowledge, only a few studies have used these assessments empirically to examine foreign language proficiency development in an online learning course. To assess differential oral proficiency outcomes for learners in face-to-face, hybrid, and online learning courses, Blake et al. (2008) employed the Versant test and found no differences in learner proficiency among the three groups, suggesting that the lack of a significant difference between each group is evidence that online instruction was at least as effective as face-to-face or hybrid foreign language instruction (n = 318: 233 in face-to-face context, 85 in hybrid or distance-learning contexts). However, since Blake et al.’s (2008) study entailed intact classes (each of which used a different instructional method in addition to a different instructional interface) and included neither a control group nor controls for potential moderating variables (such as starting proficiency, motivation, or years of prior study), the strength of inferences from their results is somewhat limited.

In a study of the maintenance and/or sustainment of Chinese language skills in online learning, Nielson et al. (2010) reported improvement in both reading and speaking proficiency scores among participants (n = 35), as measured by the STAMP test. The yearlong online learning course was designed specifically for novice to intermediate college-level learners who had already participated in a short-term face-to-face summer program in Chinese. Upon completion of a needs analysis, the course was designed to prepare learners to complete functional language tasks and, instead of a course textbook, incorporated genuine audio and video recordings of fluent speakers completing a range of these tasks in Chinese. The research design included proficiency pre-and posttests with a comparison group of learners who had completed the summer program but did not participate in the online learning Chinese course. Results indicated that students who completed the online course improved in both the reading and speaking skill modalities compared to the comparison group, as measured by the STAMP online assessment instrument.

In many circumstances, such as those measured by the proficiency assessments outlined above, the inclusion of objective indicators of learner outcomes is of fundamental importance to foreign language program administrators and other stakeholders, particularly in terms of providing evidence of learner proficiency gains. Although proficiency outcomes are important components of a larger online learning assessment framework and may provide instructors with useful information after such courses have concluded, other assessment instruments, such as formative assessments, performance-based language assessments, portfolio assessments, and self- and peer-assessments, are often incorporated into an assessment framework to document additional facets of learner gains throughout course participation. However, the extent to which these various assessment practices can be seamlessly integrated into an online learning assessment framework remains to be determined.

Assessment activities as part of a system

If the goal of assessment is to provide information to make various instructional decisions, and if there is a variety of instructional decisions that need to be made, it follows that any comprehensive assessment system will employ a variety of assessment activities to document learner achievement and proficiency, as well as provide global indicators of overall program effectiveness. To do this, there must be a general agreement among

---

19 However, counterclaims made by Downey et al. (2008) dispute this assertion, positing that learner performances on the Versant test are psycholinguistically valid and predictive of their actual performance in authentic situations.
stakeholders about the intended goals of the instruction and the objective indicators of performance that will be used as indicators of learner success.

One approach to investigating how tests are functioning in a given situation is through a testing-context analysis. The testing-context analysis attempts to systematically identify the characteristics of the context in which assessments will operate to help determine the relevance, utility, values implications, and social consequences of the use of particular assessments (Brown, 2008). The steps in a testing-context analysis are similar to those used in a needs analysis when developing a curriculum, and include (Brown, 2008, p. 282):

1. Defining the purpose of the test(s)
2. Delimiting the stakeholders
3. Deciding on the options (approaches, syllabus(es), and testing method(s))
4. Recognizing constraints
5. Selecting data collection procedures
6. Collecting data
7. Analyzing data
8. Interpreting results
9. Determining the impact of results on testing
10. Implementing testing procedures and decision making
11. Reporting on the testing-context analysis
12. Deciding on directions for future research

Data for such an analysis can come from a variety of sources, including existing information (e.g., records analysis, literature review), observations, interviews with stakeholders (structured and unstructured), meetings (e.g., focus groups), questionnaires, and even discourse analysis. (Detailed examples of most of these techniques, in the context of curriculum action research, can be found in McKernan, 1996.) Because they are already familiar with the context of the program, Bachman (1981) suggests that, despite the potential appearance of bias, formative program evaluation is best done by program staff themselves. An important first step for LanguageNation will be to determine what, if any, information can be captured through the platform itself to provide the type of information that would be useful in investigating the testing-context of LanguageNation.

Because the purposes of assessment in most programs are many, a range of assessments tools need to be employed. Although aimed more at the K-12 educational arena, Gottlieb and Nguyen (2007) offer detailed advice on how to systematically create an assessment framework for improved outcomes and accountability in a foreign language program using both formative and summative assessment tools.

One can imagine an integration of summative and formative assessment practices as operating within different instructional segments. In Figure 2 below, the large circle on the right represents the assessment of the final learning objectives for a particular course. The arrow from the rightmost circle to the leftmost circle indicates that the objectives of the course form the basis of a pre-test assessment. The blue boxes represent individual assessment tasks during the course (e.g., several course sessions dealing with a particular topic) within the larger structure, which are also related to the final objectives. Two of those individual course sessions are highlighted to show that, in a similar fashion, each of those course sessions has some instructional goal, with individual tasks geared towards that goal. In the case of LanguageNation, we can imagine a collection of individual subtasks for each of the main tasks, with specific subtasks being adaptively delivered by the system based on learner performance; stronger learners would need fewer of these subtasks. Formative assessment activities and summative activities work together to provide a coherent educational experience for the learner. The bulk of the instructional assessment would be formative in nature, with the intent of ensuring progress towards the ultimate learning objectives.
Figure 2. Course goals (circles) drive formative instructional activities (squares) and inform testing practices

**Self-assessment**

Ultimately, the learner is at the center of his or her own language development. The goal of formative assessment, discussed earlier, is to provide feedback to the learner. Encouraging learners to reflect on their own developing language through self-assessment can be especially useful when the instruction is primarily or exclusively individual. In foreign language learning, one particularly useful type of self-assessment is the “Can Do” statement. These statements encourage the learner to evaluate what he or she can do with the language in functional terms, and received widespread use as a part of the European Language Portfolio (ELP) project.\(^{20}\)

The ILR has developed a list of Can Do statements\(^ {21}\) related to the ILR skill level descriptors (e.g., “I can buy a needed item such as bus or train ticket, groceries, or clothing.”). A similar list based on the ACTFL guidelines has been developed through the LinguaFolio® project,\(^ {22}\) which is modeled on the European Language Portfolio. Can Do statements can be a useful tool for communicating performance goals. Learners can be given statements that reflect the learning objectives of the particular unit or activity that they are about to undertake, and track their progress by referring to them. In other words, the instructional objectives relevant for small sections of the course (represented by the circles in Figure 1 above) could be expressed as “Can Do”-type statements to allow for learner self-assessment before and during those units.

LinguaFolio Online (LFO), is a web-delivered self-assessment system designed to monitor changes in learners’ confidence in their abilities to perform a variety of language tasks, can be used by instructors to gauge learners’ initial estimates of their own language learning abilities.\(^ {23}\) LFO is available free of charge and comprises three main subsections. The biography component allows learners to provide information concerning their backgrounds and foreign language learning experiences. The dossier component of LFO serves as an online archive for students to upload samples of their work, such as writing samples or audio/video files of language use. Finally, the passport section provides a space for learners to upload awards, certificates, or self-reported test scores. The main component of LFO, most relevant to DA and influenced by the CEFR scale, is the self-assessment grid in the biography component of LFO. This section requires learners to assess their own

\(^{20}\) [http://www.coe.int/t/dg4/education/elp/](http://www.coe.int/t/dg4/education/elp/)


\(^{23}\) It should be noted that the paper-and-pencil version of LinguaFolio was developed by the National Council of State Supervisors for Foreign Languages (NCSSFL). The version was converted to an online version (LinguaFolio Online) by the University of Oregon Center for Applied Second Language Studies. For a complete description of LFO online, please reference [http://lfonetwork.uoregon.edu/learners](http://lfonetwork.uoregon.edu/learners).
abilities on a wide variety of language learning tasks (e.g., I can understand and say numbers in simple situations; I can ask and answer questions on familiar topics; I can interact with others using simple language in everyday situations) relative to a three-point Likert scale, ranging from (1) This is a Goal, through (2) I Can Do With Help, to (3) I Can Do. Learners are then asked to upload evidence of successful task completion, such as a video clips showcasing a successful interaction with a fluent interlocutor. Since LFO is designed entirely as a learner-centered self-assessment tool, feedback is not given to users concerning how “Can Do” statements may relate to global proficiency levels, or strategies that learners might use to progress through the Can Do scale.

Although the development of online self-assessment instruments is critical, few empirical studies have compared self-assessment instruments with outcomes on objective assessment instruments. In their two-year triangulated study designed to compare students’ self-assessments, teachers’ assessments of students, and objective measures of language proficiency in Arabic, Chinese, and Hindi (n = 748), Ross, Masters, Malone, Janesh, and Riestenberg (2010) and Masters and Ross (2011) found that teachers had more confidence in their students’ foreign language abilities than the students themselves. Objective measures of these same learners’ proficiency, measured by the Computerized Assessment of Proficiency (CAP) test (see Table 6, below), were found to fall reliably between both the students’ self-assessments of their own abilities, as measured by a subset of Can Do statements contained within the LFO, and teachers’ estimations of language learning abilities, based on the same subset of LFO Can Do statements. The researchers suggest that self-assessment accuracy perhaps could be improved by providing more comprehensive training on use of the LFO tool, offered either in face-to-face or online contexts.

**Performance-based language assessment**

In addition to self-assessment activities, such as through the use of Can Do statements, it is also possible to incorporate more formal – though not necessarily graded – assessments of functional course objectives through the use of performance based activities and scoring rubrics. Because the ultimate goal of instruction is to improve the learners’ real world capabilities with the language, it follows that there should be an emphasis on assessing learning language on tasks that represent actual language use. Assessing performance via explicit rubrics rather than merely giving generic feedback (e.g., “Nice job!”) has the advantage of being able to provide scaffolding to learners about what constitutes successful performance. To the extent performance activities reflect reasonable approximations of eventual target language tasks that the learners will need to perform, these activities can further strengthen the link between online instruction and eventual language use. In programs with multiple students, peers can assess each other using the same rubric as the instructor. Alternatively, learners can record their performances and have them viewed asynchronously by instructors or peers. And, of course, performance assessments can be used in conjunction with Can Do type self-assessment – learners can self-assess their ability to perform the tasks and then see how that self-assessment matches up with their actual performance.

For courses in which foreign language interaction is a primary course goal, effective online courses should incorporate a wide array of tools designed for learners to practice interactions with instructors, peers, and members of the target language community through text, audio, and video platforms, contextualized within a virtual environment (Blake 2009; Chapelle, 2009). To investigate the extent to which learners are subsequently able to accomplish tasks independently, criterion-referenced performance-based language assessments (PBLAs) are designed, usually by instructors or task developers, with the goal of both evaluating students’ performances and drawing inferences concerning learners’ foreign language abilities post-instruction. PBLAs typically require learners to do a variety of simulated target language tasks by engaging in meaningful interactions with instructors, conversation partners, or peers within online contexts specifically designed for foreign language learning (Chapelle, 2009).

SLA research has identified several task features that may influence learner performance on PBLAs (Bachman, 2002b; Brindley & Slatyer, 2002; Byrnes, 2002; Long & Norris, 2000; Nielson, in press; Norris, 2002; Norris et al., 2002; Robinson, 2007; Robinson & Ross, 1996; Skehan, 1998). Some task factors place cognitive or interactional demands on learners, each of which are hypothesized to affect the usefulness of task completion for language learning either positively or negatively. These factors can be manipulated by instructors or software developers. Robinson (2007) argues that three main categories should be considered when classifying L2 pedagogic tasks: (1) cognitive factors, (2) interactive factors, and (3) learner factors, each of
which will be discussed in turn. Cognitive factors are task characteristics that may influence a learner’s overall processing capacity, such as attention, memory, or reasoning, and are manipulated directly by the cognitive demands of task completion; these may change from moment to moment, relative to task design. He argues that the overall category of cognitive factors can be further divided into two subcategories: (a) resource-directing variables, such as here-and-now versus there-and-then task designs, and (b) resource-dispersing variables, such as the provision or lack of planning time (p. 17) prior to task completion. Interactive factors are defined as those that influence the conditions within which task completion takes place. Robinson also divides the overall category of interactive factors into two main subcategories: (a) participation variables, such as task conditions in which there are open versus closed solutions, and (b) participant variables, such as the gender or proficiency level of interlocutors involved in task completion. Finally, Robinson defines learner factors as the individual difference variables inherent to each learner that cannot change from moment to moment. Consistent with his description of cognitive and interactive factors, Robinson further divides learner factors into two subcategories: (a) individual learner ability differences, such as high or low working memory or aptitude, and (b) affective variables, such as high or low motivation or anxiety.

While it is likely that the pedagogic L2 task classifications outlined by Robinson influence task completion, virtually no empirical research has investigated the unique influence that task characteristics such as planning time (Skehan, 1998), abstractness (Robinson, 1995), solution type (Long, 1989), and familiarity (Plough & Gass, 1993) may have on learners specifically within the computer-mediated context. Furthermore, no empirical research has examined the relationship between the format of task conditions (i.e., visual, audio, video) or temporal conditions (i.e., synchronous or asynchronous) and learners’ abilities to engage in simulated interactions with target language interlocutors within the virtual environment. (For a detailed breakdown of L2 pedagogic task classifications and their associated subcategories, see Appendix B.)

Since it is likely that these factors may also influence inferences made concerning learners’ abilities to complete simulated target language tasks within the virtual environment, the design and development of valid and reliable PBLA instruments requires program administrators and instructors to decide on and document, from the beginning, those features of language use that will count as evidence of successful task performance, as well as the criterion-referenced rating scales to be used in evaluating learner performance (Bachman & Palmer, 1996; Mislevy et al., 2002). To this end, it is essential for PBLA designers to carefully construct tasks that appropriately elicit the desired features of linguistic performance from learners, as well as specify the environmental conditions and contexts within which the various tasks will be performed.

While such documentation is essential for valid and reliable PBLAs, few empirical studies have reported outcomes on task-based language performance assessments in online learning contexts. Perhaps this can be attributed to the technical challenges associated with systematically capturing evidence of learner performances within an online environment. In reporting PBLA outcomes in a prototype online Chinese course, Nielson (in press) employed Rasch analysis to examine the extent to which task item difficulty estimates approximated the range of learner ability estimates. At the conclusion of each task-based module designed around such target tasks as ordering food in a restaurant, buying a cell phone, or arranging a trip, students were asked to complete PBLAs by role-playing each module’s target task with a fluent conversation partner, using VoIP technology. Results identified two subtasks that violated the probabilistic model, indicating a potential disconnect between the type of behavior intended to be elicited from the assessment rubric and the actual linguistic behavior elicited from the task design.

Although challenging, it is imperative for PBLA designers to implement systematic, ongoing validation procedures for all PBLA instruments employed within an online learning setting, especially since courses may require learners to exploit the technology to engage in a variety of authentic tasks. Since PBLAs often generate complex linguistic data from learners, for PBLA designers to make systematic progress in creating both valid and reliable instruments for use within a performance-based online setting, the specification of task conditions within which the performance-based assessments will take place is of utmost importance. This includes the unique features of task performance within the online context, as well as the documentation of what counts as

24 Rasch measurement is a sample-free item response theory (IRT) form of measurement that places person ability estimates and item difficulty estimates on an equal interval, probabilistic scale.

25 ‘Subtasks’ refer to subcomponents of task completion that together account for successful task performance.
evidence of learner success. Furthermore, the incorporation of Rasch modeling to identify misfitting performance tasks and/or examine the impact of rater severity, as well as the appropriateness of rating scales in evaluating learner performances, can provide assessment designers with the objective information necessary to make evidence-based adjustments to the design of PBLA instruments.

CHALLENGES AND OPPORTUNITIES

The computer-mediated format of online language courses creates some challenges for assessment, but perhaps the biggest challenge is not unique to the online environment. At present, most online testing is not so radically different from the type of testing that has been done for decades in more traditional settings. And, although computer-based testing is rapidly becoming commonplace even in face-to-face instructional programs, those computer-delivered tests are still reminiscent of the paper-and-pencil tests they were intended to replace rather than representing a radically new assessment paradigm (Chalhoub-Deville, 2001). For many instructors, especially in the age of accountability, assessment is still generally seen as an imposed burden that interferes with instruction rather than a potentially rich source of information that enhances it.

Even though there has been much interest in formative assessment since the publication of Black and Wiliam’s (1998) seminal article documenting its usefulness, this has not necessarily resulted in changes to classroom practice, online or otherwise. Many pre-service instructors still think of assessments from a summative perspective (Volante & Fazio, 2007). Although teacher-created assessments and other tools such as portfolios and self-assessments are frequently cited as examples of formative assessment, it is the feedback aspect of assessment that makes it formative, not the format of the assessment tool itself. Not all multiple-choice tests are summative, and not all performance-based tests are formative. Effective teachers have always intuitively provided feedback to learners, though perhaps not in as systematic a fashion as possible and, as Harlen and James (1997) have noted, what many instructors implement in the name of formative assessment is actually just a series of small summative assessments.

Although the project suffered funding cuts before coming to fruition, Clark (2011) proposed the development of an online assessment delivery platform that would allow users to create assessment instruments by drawing on an instructor-developed bank of assessment items. Similar to the crowdsourcing of encyclopedia entries pioneered by Wikipedia, the vision was to encourage instructors, especially those of underserved languages, to create and share assessment items through an online, searchable repository. Instructor-developed assessment tasks, including associated scoring rubrics for performance items, would be available for the mass customization of assessment instruments. For performance based items, rater training using actual student samples and anonymous grading queues could be integrated into the system. The SLUPE project\textsuperscript{26} represents a similar crowdsourced approach applied to language placement.

The LanguageNation platform has the potential to change this reality as we finally begin to incorporate technology for training and testing in cutting edge ways. There have been calls in the general educational assessment literature to base assessment instruments on statistical models of student learning (Pelligrino, Chudowski, & Glaser, 2001). These student models would contain updated probabilities about student performance, accumulated across assessment tasks. To the extent that these statistical models are underpinned by models of learner cognition or development, there is the potential to provide more sensitive diagnostic information about learning trajectories. Graphical “construct maps” that depict a learner’s position on a continuum of ability, from beginning to proficient, can be developed (Wilson, 2005). Suggestions of how some of these modeling ideas might be applied to task-based language testing are given in Mislevy, Steinberg, and Almond (2002).

Future innovations could make possible continuous assessment, in which there is no single assessment event \textit{per se}, but rather an ongoing evaluation of learning. Because the computer is capable of tracking learner interaction over time, there is the potential for the development of a student model in which all of the online learning experiences of the student are aggregated into an ever-evolving model of student performance. As more

\textsuperscript{26} \url{http://phrants.net/pt/pt.html}
and more applications become “cloud based” rather than residing on the particular learner’s computing device, it is possible, in theory, to track all of the target language interactions that a learner has. This possibility has led to work on specifications for tracking learning experiences across different online learning platforms.27 If these innovations are successful and resources are built that incorporate such standards, it could be possible, for example, to track the time spent and the number of questions correctly answered by a learner in a GLOSS module that was done for homework. A similar process would happen for the learner’s other online activities. The instructor would be able to review this information to make pedagogical recommendations. Individually, this information is only minimally informative. However, it might be possible to estimate a learner’s developing proficiency by aggregating these results over many tasks. In other words, rather than being the product of a single test session on a single day, the state of a learner’s ability would be constantly updated as it is assessed across a variety of tasks over time. The ultimate goal is not merely the aggregation of individual performances, but rather the collection of evidence that reveals underlying ability.

**Recommendations for LanguageNation assessment**

As stated early on in this report, the goal of assessment is to gather information to make some kind of decision. Because the types of decisions needed in LanguageNation will evolve as the project moves forward, ongoing investigation of assessment issues will be needed. There is no such thing as the “right” test; different situations may require different approaches. The following list is a combination of recommendations for practice as well as potentially useful avenues of exploration as the LanguageNation platform evolves.

- Prioritize the quality of information over technical expediency
- Collect information from a variety of sources
- Explore crowd-sourced assessment task development
- Provide feedback on progress to the learners
- Encourage learner self-reflection on progress
- Investigate the use of construct maps (Wilson, 2005) to visualize learner progress
- Align summative assessments with course objectives
- Recognize that comparison to native-speaker norms may not be appropriate in the early stages of language development
- Explore the feasibility of mimicking the adaptivity of the oral proficiency interview
- Investigate the use of task-based assessments
- Employ multiple measures of learner progress

---

REFERENCES


Clifford, R. (2009, April). When what you have is not enough. Workshop conducted at National Council of Less Commonly Taught Languages (NCOCTL), Madison, WI.


© 2013 University of Maryland. All rights reserved.

UNCLASSIFIED


# APPENDIX A (ADAPTED FROM BENSON ET AL., 2012)

General description of some commercially available, online proficiency tests

<table>
<thead>
<tr>
<th>Assessment name</th>
<th>Languages offered</th>
<th>Skills assessed</th>
<th>Test characteristics</th>
<th>Factors impacting relevance to online learning</th>
<th>Website Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>Online Diagnostic Assessment (ODA)</td>
<td>Arabic (MSA) Chinese Farsi Korean Portuguese (Brazilian) Russian Spanish Tagalog</td>
<td>Reading comprehension Listening comprehension</td>
<td>Web-based, semi-adaptive, automatically scored 1 to 2 hour duration Employs authentic texts</td>
<td>Free of charge Format comparable to DLPT Feedback given in form of diagnostic profile, not ACTFL/ILR levels</td>
<td>ODA</td>
</tr>
<tr>
<td>Computerized Assessment of Proficiency (CAP)</td>
<td>Arabic (MSA) Chinese Hebrew Hindi Persian Swahili Turkish Urdu Yoruba</td>
<td>Presentational writing Presentational speaking Contextualized grammar Interpretive listening Interpretive speaking</td>
<td>Free of charge to language teachers Pilot test of future STAMP (below) items</td>
<td>Scoring presentational speaking, writing requires trained raters not provided by CAP</td>
<td>CAP</td>
</tr>
<tr>
<td>Standards-based Measurement of Proficiency (STAMP)</td>
<td>Arabic (MSA) Chinese English French German Japanese Spanish</td>
<td>Presentational writing Presentational speaking Interpretive listening Interpretive speaking</td>
<td>Web-based Employs ACTFL proficiency guidelines</td>
<td>Requires physical oversight by proctor</td>
<td>AVANT - STAMP</td>
</tr>
<tr>
<td>ACTFL Assessment of Performance toward Proficiency in Languages (AAPPL)</td>
<td>Arabic (MSA) Chinese French German Russian Spanish</td>
<td>Interpersonal speaking Interpretative reading Interpretive listening Presentational writing</td>
<td>Web-based Administered via a video virtual instructor Tasks include video chat, creating wikis, writing online journal entries, etc. in target language</td>
<td>Differences between individual curricula and AAPPL assessment tasks may lead to inaccurate proficiency assessments</td>
<td>AAPPL AAPPL - Video</td>
</tr>
<tr>
<td>Computerized Oral Proficiency Instrument (COPI)</td>
<td>Arabic (MSA) Spanish</td>
<td>Overall oral proficiency</td>
<td>Computer-based Semi-adaptive Employs ACTFL proficiency guidelines</td>
<td>Speaking scoring requires trained raters</td>
<td>COPI</td>
</tr>
<tr>
<td>Assessment name</td>
<td>Languages offered</td>
<td>Skills assessed</td>
<td>Test characteristics</td>
<td>Factors impacting relevance to online learning</td>
<td>Website Link</td>
</tr>
<tr>
<td>---------------------------------------------------</td>
<td>------------------------------------------</td>
<td>---------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------</td>
<td>------------------------------------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Versant Automated Oral Proficiency Tests</td>
<td>Arabic (MSA) French Spanish</td>
<td>Overall oral proficiency, separate scores for sentence mastery, vocabulary, fluency and pronunciation</td>
<td>Web- or phone-based Test tasks include sentence repetition, responding to rearranging groups of words to produce sentences, retelling passages Scored using ASR software Employs ACTFL proficiency guidelines</td>
<td>Use of ASR constrains task type Inauthentic, decontextualized testing interface Short duration (15–17 minutes) Easy to administer</td>
<td>Versant</td>
</tr>
<tr>
<td>ALTA language assessments</td>
<td>More than 80 languages and dialects</td>
<td>Speaking and listening tests, online reading assessments, and proctored writing assessments</td>
<td>Phone-based Listening and speaking assessed according to speed, range of comprehension, structure complexity, accent, accuracy of verb use, vocabulary range</td>
<td>Assessment criteria vague and subjective No information on ALTA score correspondence to ACTFL/ILR guidelines</td>
<td>ALTA Language Assessments</td>
</tr>
<tr>
<td>The Minnesota Language Proficiency Assessments (MLPA)</td>
<td>French German Spanish</td>
<td>Contextualized reading Contextualized writing Contextualized listening Contextualized speaking</td>
<td>Reading, writing, listening, speaking test Immediate automated scoring on reading, listening; writing and speaking scored by trained raters Employs ACTFL proficiency guidelines</td>
<td>Online writing and speaking requires identification of trained raters Rubric and rating scale not available for review on website</td>
<td>MLPA</td>
</tr>
</tbody>
</table>

28 At http://www.pearsonhighered.com/Versant, the Versant test is referred to as a proficiency test; however, at http://versanttest.com/products.jsp, it is referred to as a language test.
<table>
<thead>
<tr>
<th>Assessment name</th>
<th>Languages offered</th>
<th>Skills assessed</th>
<th>Test characteristics</th>
<th>Factors impacting relevance to online learning</th>
<th>Website Link</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>The Very Low Range Defense Language Proficiency Tests</strong></td>
<td>Arabic (Iraqi) Chinese (Mandarin) Dari French Korean Pashto Persian-Farsi Spanish Urdu</td>
<td>Reading comprehension Listening comprehension</td>
<td>90 minutes for each of two sections Targets learners in the 0+ – 1+ range All questions in English Employs ILR proficiency guidelines</td>
<td>Relatively new, untested Score may depend on lexical items previously studied and not overall proficiency</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>American Council on The Teaching of Foreign Languages Internet-delivered Writing Proficiency Test (ACTFL iWPT)</strong></td>
<td>Albanian Arabic (MSA) Chinese (Cantonese, Mandarin) Dutch English German Greek Hebrew Hindi Italian Japanese Korean Persian Polish Portuguese Russian Serbo-Croatian Spanish Turkish Urdu Vietnamese</td>
<td>Writing</td>
<td>Web-based Employs ACTFL proficiency guidelines Rated by ACTFL Certified WPT Raters 40-80 minutes depending on the level of the candidate</td>
<td>Requires physical oversight by proctor</td>
<td><a href="#">ACTFL iWPT</a></td>
</tr>
<tr>
<td>Assessment name</td>
<td>Languages offered</td>
<td>Skills assessed</td>
<td>Test characteristics</td>
<td>Factors impacting relevance to online learning</td>
<td>Website Link</td>
</tr>
<tr>
<td>--------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------</td>
<td>-----------------</td>
<td>---------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>American Council on The Teaching of Foreign Languages Oral Proficiency Interview by Computer (ACTFL OPIc)</td>
<td>Afrikaans (MSA) Bengali Chinese (Mandarin) English French German Indonesian Korean Pashto Persian Russian Spanish Tagalog</td>
<td>Speaking</td>
<td>Web-based Rated by ACTFL certified ILR OPIc raters</td>
<td>Short duration (approximately 30 min) Scores only go up to ILR level 2</td>
<td>ACTFL OPIc</td>
</tr>
<tr>
<td>Computer Assisted Screening Tool (CAST)</td>
<td>Arabic (MSA, Egyptian) Chinese (Mandarin) Dari English Farsi/Persian Filipino French German Hindi Italian Japanese Korean Pashto Russian Spanish</td>
<td>Speaking</td>
<td>Web-based Rated by CAST assessor</td>
<td>Short duration (15 to 30 min) Feedback aligned with ACTFL and ILR oral proficiency guidelines Useful only for intermediate and advanced levels</td>
<td>CAST</td>
</tr>
</tbody>
</table>
APPENDIX B. Pedagogic L2 Task Classification—Categories, Criteria, Analytic Procedures, and Characteristics (adapted from Robinson, 2007, p. 15)

<table>
<thead>
<tr>
<th>Task Complexity</th>
<th>Task Condition</th>
<th>Task Difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive Factors</td>
<td>Interactive Factors</td>
<td>Learner Factors</td>
</tr>
<tr>
<td>Classification criteria: cognitive demands</td>
<td>Classification criteria: interactional demands</td>
<td>Classification criteria: ability requirements</td>
</tr>
<tr>
<td>Classification procedure: information-theoretic analyses</td>
<td>Classification procedure: behavior descriptive analyses</td>
<td>Classification procedure: ability assessment analyses</td>
</tr>
<tr>
<td>Subcategory (a) Resource-directing variables making cognitive/conceptual demands</td>
<td>Subcategory (a) Participation variables making interactional demands</td>
<td>Subcategory (a) Ability variables and task-relevant resource differentials</td>
</tr>
<tr>
<td>+/- here and now (Robinson, 1995)</td>
<td>+/- open solution (Long, 1989)</td>
<td>h/l working memory (Mackey et al., 2002)</td>
</tr>
<tr>
<td>+/- few elements (Kuiken et al., 2005)</td>
<td>+/- one-way flow (Pica et al., 1993)</td>
<td>h/l reasoning (Stanovitch, 1999)</td>
</tr>
<tr>
<td>+/- spatial reasoning (Becker &amp; Carroll, 1997)</td>
<td>+/- convergent solution (Duff, 1986)</td>
<td>h/l task switching (Monsell, 2003)</td>
</tr>
<tr>
<td>+/- causal reasoning (Robinson, 2005a)</td>
<td>+/- few participants (Crookes, 1986)</td>
<td>h/l aptitude (Robinson, 2005b)</td>
</tr>
<tr>
<td>+/- interactional reasoning (Baron-Cohen, 1995)</td>
<td>+/- few contributions needed (McGrath, 1984)</td>
<td>h/l field independence (Skehan, 1998)</td>
</tr>
<tr>
<td>+/- perspective taking (MacWhinney, 1999)</td>
<td>+/- negotiation not needed (Long, 1983)</td>
<td>h/l mind-reading (Langdon et al., 2002)</td>
</tr>
<tr>
<td>Subcategory (b) Resource-dispersing variables making performative/procedural demands</td>
<td>Subcategory (b) Participant variables making interactant demands</td>
<td>Subcategory (b) Affective variables and task-relevant state-trait differentials</td>
</tr>
<tr>
<td>+/- planning time (Skehan, 1998)</td>
<td>+/- same proficiency (Yule &amp; MacDonald, 1990)</td>
<td>h/l openness (Costa &amp; Macrae, 1985)</td>
</tr>
<tr>
<td>+/- prior knowledge (Urwin, 1999)</td>
<td>+/- same gender (Pica et al., 1991)</td>
<td>h/l control of emotion (Mayer et al., 2002)</td>
</tr>
<tr>
<td>+/- single task (Robinson et al., 1995)</td>
<td>+/- familiar (Plough &amp; Gass, 1993)</td>
<td>h/l task motivation (Dörnyei, 2002)</td>
</tr>
<tr>
<td>+/- task structure (Skehan &amp; Foster, 1999)</td>
<td>+/- equal status/role (Yule &amp; MacDonald, 1990)</td>
<td>h/l processing anxiety (MacIntyre &amp; Gardner, 1994)</td>
</tr>
<tr>
<td>+/- few steps (Fleishman &amp; Quaintance, 1984)</td>
<td>+/- shared cultural knowledge (Brindley, 1987)</td>
<td>h/l willingness to communicate (MacIntyre, 2002)</td>
</tr>
<tr>
<td>+/- independency of steps (Romiszowski, 1988)</td>
<td></td>
<td>h/l self-efficacy (Bandura, 1997)</td>
</tr>
</tbody>
</table>

Note: +/- denotes presence or absence of demands during task completion; h/l denotes high or low ability brought to the task by the learner.
APPENDIX C – Review of available assessments in Persian Farsi/Dari and Somali

CASL researchers with experience in the initial LanguageNation languages were asked to identify and comment on existing assessments in those language.

DLIFLC’s Online Diagnostic Assessment System (ODA)

Web Address: http://oda.lingnet.org/

Languages: Persian Farsi. Dari is also listed on the site, but currently the Reading Assessment is listed as “in validation,” while the Listening Assessment is “under development.” Somali is “under development”.

Access: can be accessed by the public. New users can register by visiting the address above, clicking the “Log In” tab, and clicking the “Register” button.

Persian Farsi Reviewer Comments: the assessments consist of relatively lengthy authentic texts or audio clips, followed by comprehension questions. In terms of format, there are multiple choice questions where the user must identify the correct textual quote/audio clip, type in the English translation of selected terms and phrases, or identify correct areas within the Persian text (e.g. “Select the object of this verb in the text.”).

The breakdown provided upon completing an assessment is fairly detailed. It provides an estimated ILR level based on the user’s performance, and goes on to score specific areas such as comprehension of main ideas and supporting ideas, vocabulary, performance within the subject areas of Geography, Military, and Culture, comprehension of specific structural (grammatical) elements, etc.

Appropriate for: technically, users of any ILR level. The ODA is:

…semi-adaptive, which means the number of passages and questions you get will depend on your performance. If you perform well at an advanced level of content, you will most likely go through more passages and questions than someone who performs well at an intermediate or lower level. At the very least, you will go through three sets of passages (for a total of nine passages).

The system does have a limited ability to “ratchet down” the assessment for lower-level users, but this has a more pronounced effect on the length of the assessment rather than on the complexity of the texts and clips used. In other words, the system will not provide elementary texts or clips for users that are performing very poorly.
Joint Language University (JLU)

Web Address: http://jlu.wbtrain.com/sumtotal/jlu2.0/HOME/index.asp

Languages: Persian Farsi, Dari, Somali.

Access: users can register at https://www.wbtrain.com/SumTotal/JLU2.0/registration/index.asp, however, they are asked to select which organization they are affiliated with from a list that consists of government agencies, the military, and a few foreign governments (the U.K., New Zealand, Australia, and Canada).

Persian Farsi/Dari Reviewer Comments: the online catalog contains sections for Area & Culture Training, Language Training, and Web Based Assessments. Within Web Based Assessments (sic), there are sections for Reading Skills and Listening Skills (there are other sections such as General Language that currently contain no items).

The reading and listening assessments consist of comprehension questions about authentic texts and audio clips. The assessments are self-guided, and only provide students with feedback on whether or not their answers were correct; they are not provided with a plan or milestones for increasing their level.

The quality of the audio clips is variable. Many of them contain relatively high levels of white noise, or are somewhat faint.

The description of each assessment contains a recommendation for what level the user should be in order to use it (e.g. “The passages and questions are appropriate for ILR proficiency level 3.”) The assessments score the user’s answers, but don’t diagnose their level. In fact, the description of each assessment contains a disclaimer such as “Successfully completing this Assessment Object does not guarantee that you can perform consistently at proficiency level 3.”

Appropriate for: students between ILR levels 1+ and 3. The inventory of assessments includes:

- Dari: 39 assessments at 1+/2, 40 assessments at 2+/3.
- Persian Farsi: 20 assessments at 1+/2, 20 assessments at 2+/3.

Somali Reviewer Comments:

- The modules contained within JLU appear to be a superset of those found via DLIFLC. Within the language training section, there are self-paced and instructor-led training tabs. For Somali, there are:
  - 109 total objects:
• These objects can be sorted by ILR level, topic, content provider, and knowledge type (cultural, reading, etc.)
• One module contains a stand-alone Somali Familiarization Course developed for USSOCOM. It is described as a 200 hour course designed to bring students to the 0+/1 level. It contains a Language Basics section, as well 20 lessons on a variety of topics. Some lessons rely heavily on a lesson partner to complete. Each lesson contains a number of activities, some of which have accompanying sound clips to demonstrate pronunciation. The “End of Lessons Tasks” are static and rely on the student to compare their own work to samples in the answer key. No feedback is provided. A static vocabulary list without example clips is provided.
• Listening exercises contain transcripts and translations and a few example words in a glossary tab.
• Besides this Familiarization Course, there is no instruction in the Self-Paced training modules. There are no consistently offered instructor-led training courses in Somali.
• There are no web-based assessments for Somali.
• There is no Rosetta Stone module for Somali.
• CL-150 Transparent Language has been disabled due to funding shortfalls.
• Previously able to find through JLU a more extensive, modernized USSOCOM 200-Hour Language Familiarization course for Somali. Includes 247 page ‘textbook’ to accompany online learning materials. Comes with listening exercises. No grammar explanation. There is also a 73 page Instructor Handbook, though it is not language-specific like the textbook. This appears to be the most complete online learning Somali course. Language taught in chunks, but content is sometimes misleading. This is no longer available via JLU.
Headstart2

Web Address: http://hs2.lingnet.org/

Languages: Persian Farsi, Dari, Somali.

Access: users can register by clicking at the web address listed above. They are not required to state an affiliation with any organization. The site does request access to the user’s camera and microphone, and says that users who grant access may be recorded.

Persian Farsi/Dari Reviewer Comments: users answer questions in each module to determine whether or not they mastered the language skills and vocabulary in that module. However, the assessment is restricted to “GO” or “NO GO” (pass/fail). The student’s skills are not analyzed, and they are not given a tailored plan to address their personal problem areas (other than to presumably do and repeat all of the modules if necessary).

The courses seemed to be designed to rapidly train users in a very limited vocabulary set of particular use in military situations. This manifests itself in examples that include but are by no means limited to:

- Later modules require users to select the correct translation for phrases such as ﻣﺎ دﻧﺒﺎل مواد ﺑﻤﺐ ﺳﺎزی ﻫﺴﺘﯿﻢ (“We are searching for bomb-making materials.”).
- Grammatical instruction is deliberately limited; for example users are only taught how to give basic, informal commands in the ﺗﻮ form ( ﺗﻮ is similar to tú in Spanish and other Romance languages).
- There are virtual characters that pronounce key phrases and words, and move their mouths in order to show the approximate mouth shape needed to pronounce the items in question correctly. Many of the characters are dressed as U.S. military personnel.
- Simplified transcription of key words and phrases is provided even up through the later modules, e.g. “motaRjem kojaas?” for ﻣﺘﺮﺟﻢ ﮐﺠﺎس؟
- The major differences between formal and colloquial Persian are de-emphasized, or perhaps not addressed at all.

Appropriate for: military personnel who need to rapidly attain elementary proficiency in Persian Farsi or Dari in order to meet basic daily needs and handle a limited set of critical situations.

Headstart2 could hypothetically be used by a new, non-military learner of Persian, but he or she would probably have to move on from Headstart2 to a more advanced curriculum within a few weeks to a month (approximately).

Somali Reviewer Comments:

- Recently released module for Somali, available Online, IPod, and download with free registration, i.e. open access
Two, ten-unit modules (Sound & Script, Military)

There is a sound recorder tool, which I believe is intended for testing comparison between spoken sound clip examples and the student’s own pronunciation. The problem, however, is that it relies on the student’s impressionistic comparison with no feedback. Also in terms of sound is a ‘Somali pronunciation guide’ which provides sound clips of 5 sounds that may give native English speakers difficulty.

Glossary module is only partly complete, but contains phrases, rather than words. For each, there is an accompanying sound file. English gloss, Somali transliteration, and Somali gloss are given. This is more of a reference tool than a training tool.

Modules contain a variety of activities (Steps) built around a Task. Each Step has a Purpose and Directions drop-down tab; however each module is centered around rote learning. For most Steps, a Grammar Note and Cultural Note are provided; however these do not often relate directly to the Tasks.

There is no instruction on vocabulary; instead the modules rely on rote learning to complete tasks. Evaluation is based on whether or not the tasks are completed successfully, but no substantive feedback is provided.

- For example, in Step 4 of Task 1, a tic-tac-toe game is given. Students are to listen to sound clips and identify words. While the sound identification component is helpful, feedback is inconsistent, with some grammar and some cultural notes given.
- Module 2, Step 1 relates to numbers. The first task is a spelling fill in the blank; however no tutorial on numbers is given, so the student has no prior instruction to build upon.
- Steps called “Introduction to Vocabulary” rely on students recognizing cartoon drawings and fill in the blank for listening practice; however when a student enters an incorrect response, the correct spelling first given, but then the student is given a second chance to try the item.

Grammar notes do not immediately correspond to items taught in the Modules.

Global Online Language Support System (GLOSS)

Web Address: http://gloss.dlifc.edu/Default.aspx

Languages: Persian Farsi, Dari, Somali.

Access: users can access the lessons by visiting the address above, with no registration required.
Persian Farsi/Dari Reviewer Comments: there is no assessment on this webpage per se. When one clicks on the “Diagnostic Assessment” tab, they are directed to the DLI’s ODA.

Like material available on JLU, the material on GLOSS consists of authentic texts and audio clips that users must examine, and then answer comprehension questions. The user’s answers are scored, but the user doesn’t receive a diagnosis of their current ILR level. In fact, the user accesses material by specifically selecting the level of material they want to see.

The main page states:

In-depth Feedback accompanies all motivating tasks, providing learners with thorough explanations and tutoring just like an attentive and experienced teacher would do.

However, this in-depth feedback seems to consist of built-in textual content, rather than any kind of system that analyzes and provides strategies for addressing the user’s problems areas. Some of the audio content suffers from the same lack of quality as the material on JLU.

Appropriate for: students between ILR levels 1 and 3.

Somali Reviewer Comments:

- Contains 15 Somali lessons for reading and listening drawn from authentic materials. The lessons can be completed online or downloaded. The ‘compare’ button provides some defined vocabulary items to be encountered in the text/recording. Feedback as to which choice is correct is given, alongside a glossed translation and ‘notes’ containing defined vocabulary items.

- For reading modules, everything is by rote, and only upon finally discovering the correct response is an explanation given with a gloss and discussion of important vocabulary items. There is no chance of the student learning any productive skills from the material.

FAMiliarization –

- Contains a “Language Survival Kit” with 10 sections mostly related to military-specific terms and phrases. This provides recordings in English and Somali, include a Romanization of the Somali orthography. There is no practice or evaluation mechanism.