EMAN ELTURKI
United States

A Systematic Process for Assessing Assessment

How to assess assessment is often a dilemma for teachers and English language programs. Accrediting agencies for English language programs, such as the Commission on English Language Program Accreditation (CEA), require a plan in writing for monitoring and reviewing assessment practices. Nonetheless, web-search queries such as "assessing assessment," "how to assess assessment," "assessing assessment reliability and validity," and so on do not yield results that provide a systematic process for assessing assessments. My multiple search attempts resulted in articles related to, for example, classroom strategies for assessing students, the use of rubrics, areas to be considered when developing assessments, and explanations of assessment-related concepts such as validity and reliability. These are valuable resources for assessment but do not suffice for what I wanted: a comprehensive assessment framework that guides the phases of the assessment cycle, from pre-assessment creation to post-assessment analyses.

This article is a result of research on how to revitalize assessment practices in a university-based intensive English language program in order to maximize validity and reliability and meet accreditation requirements. This includes standardized measures and guidelines for teachers to draw upon (1) *prior* to constructing an assessment, (2) *when* constructing the assessment, (3) *while* administering the assessment, (4) *when* grading the assessment, and (5) *after* grading the assessment. Programs as well as individual teachers can adapt this assessment framework to enhance their assessment practices and fulfill accreditation requirements.

CURRICULUM–ASSESSMENT INTERDEPENDENCE

A clear course goal in line with the program's mission, specific course objectives, and measurable student learning outcomes (SLOs) are infrastructural components of a course. CEA's Curriculum Standard 2 mandates that a language program's "course goals, course objectives, and student learning outcomes are written, appropriate for the curriculum, and aligned with each other. The student learning outcomes within the curriculum represent significant progress or accomplishment" (CEA 2019). To achieve the objectives and SLOs set for a given course, educators must make sure that level-appropriate instructional materials and activities, formative and summative methods of assessment, and a grading scheme are in place. These elements of the curriculum are interconnected and must be accounted for throughout the teaching and learning cycle (see Figure 1).

Assessment is a core component of teaching and learning. Within the context of English

Programs as well as individual teachers can adapt this assessment framework to enhance their assessment practices and fulfill accreditation requirements.

as a second or foreign language (ESL/EFL) teaching, instruction focuses on developing students' language abilities, while assessment allows for gathering data on language abilities. Hence, language teaching and language assessment complement each other (Bachman and Damböck 2018). Two types of assessments are used to collect information on instruction and students' learning. The first type is *formative assessment*, which aims to evaluate students in the process of learning and forming their skills. Data collected from formative assessments help monitor learning and inform teaching. That is, formative assessment allows the course instructor to *form* an idea about his or her teaching, instructional materials, and learning tasks and provide feedback to students on learning

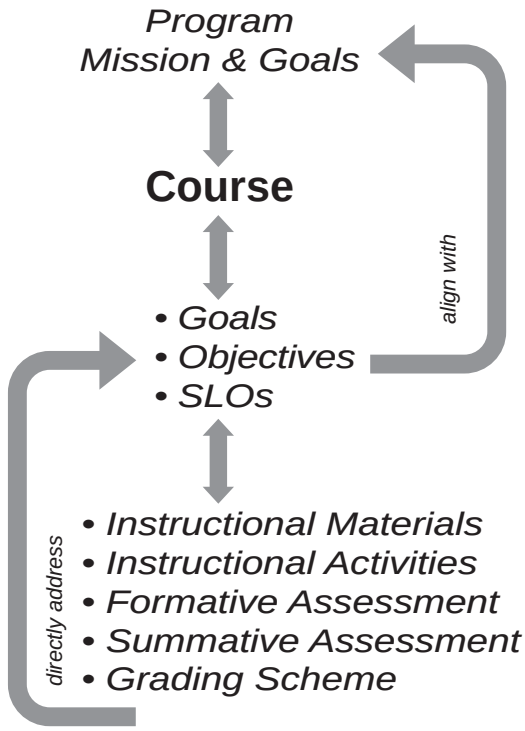

Figure 1. Curricular elements and their interconnectedness

(Bachman and Damböck 2018). Formative assessments are typically low stakes (i.e., have no or a low-grade value) and can take place prior to, during, and after teaching and learning. Examples of formative assessments include worksheets, quizzes, in-class activities, homework exercises, individual conferences, and a teacher's observations of students during group/peer work.

Summative assessment, on the other hand, *sums up* what students have learned, determines whether students reached benchmarks, and measures how well students have met the learning outcomes of a course. Summative assessments are generally high stakes (i.e., have a high-grade value) and take place at the end of a unit or course. Examples of summative assessments are end-of-unit or chapter tests, midterm and final exams or projects, and standardized tests.

PRINCIPLES OF ASSESSMENT EFFECTIVENESS

Fundamentally speaking, an effective test "is expected to yield valid and useful score-based interpretations about what the examinees know and are able to do with respect to a defined target domain" (Birenbaum 2007, 29). In language testing, Bachman and Palmer (1996) propose a model of test usefulness to inform the design of language assessment. The main premise of this model is that language assessments should yield information about a student's language abilities that can be generalized beyond the test itself. That is, language testing should resemble language that is used in a natural environment, which is in line with the principles of the communicative language teaching approach. This model of test usefulness is commonly adopted in

Assessment is a core component of teaching and learning.

language testing (Priyanti 2017; Schmidgall, Getman, and Zu 2018; Thu 2019) to assess a test's effectiveness against the following six qualities (Bachman 2001; Bachman and Palmer 1996):

1. **Reliability.** Reliability refers to the consistency of results. That is, if an exam is given at different times under similar conditions, the test should yield consistent scores. If a student takes the same grammar test at a later time (assuming similar conditions and no additional instruction) and gets a drastically different score, then the test can be considered unreliable. Reliable assessment instruments and procedures yield dependable and consistent information.

2. **Validity.** Test validity means that a test accurately measures what it is intended to measure. For instance, a speaking test should measure students' speaking ability by engaging them in actual speaking tasks. If the test instead asks students to complete written dialogues, then the test is invalid because it does not assess the skill that it purports to measure.

3. **Authenticity.** Authenticity means that test tasks are relevant and meaningful, and that they correspond to real-world contexts. Authentic assessment tasks in a course that aims to prepare students to give academic presentations at the graduate level would involve designing and delivering actual presentations to an audience (e.g., classmates and teachers) using PowerPoint slides or posters on topics related to their majors. Authentic assessments allow for interpreting and generalizing test scores beyond the testing task.

4. **Interactiveness.** Interactiveness refers to the degree to which students' individual language and cognitive abilities, personal characteristics, and background knowledge influence their performance on a test. Optimizing students' interactiveness with a given test involves considering, for example, their age and level of education in the first language. For instance, collecting information about students' level and type of education in a foundational course for adult English language learners may reveal that some students lack minimal literacy skills in their native language. Such information is valuable in test design and development.

5. **Impact.** Impact pertains to the consequences that test results might have on an individual (impact at the micro level) and the educational system or society at large (impact at the macro level). The impact of a test at the micro level can take different shapes and forms, from generating feedback on language development—which can inform teaching and learning—to enhancing educational or career opportunities. Examples of a test impact at the macro level include modifying the curriculum of a certain course in a program due to consistent high fail rates.

A valid and reliable assessment allows for making inferences about students' language abilities beyond the test itself.

6. **Practicality.** Practicality means that a test is developed and implemented within the limits of available resources. A practical test does not consume a large amount of time to develop, administer, or grade and is carried out using available materials, space, and equipment.

In summary, a valid and reliable assessment allows for making inferences about students' language abilities beyond the test itself. The proposed framework in this article offers procedures to guide the design, grading, and evaluation of an assessment. The main goal is to increase reliability and validity of language testing with the assumption that accredited English language programs do the following:

1. Adopt communicative language teaching and, as a result, develop assessment tasks that are meaningful and purposeful (authenticity).
2. Design courses that take into account students' academic and personal needs as well as linguistic and sociocultural backgrounds through, for example, conducting needs analysis (interactiveness).
3. Use assessment data to provide feedback to students on language development, improve pedagogical practices, inform curricular changes, and make decisions regarding mastery of SLOs and progression (impact).
4. Consider available resources when designing and administering an assessment task (practicality).

The proposed assessment framework functions as a guide to assessment expectations and procedures for faculty and language programs "throughout the test development process, and not simply after

Formative Assessments	Summative Assessments
• Evaluating students in the process of forming their skills to assess progress and inform teaching (e.g., quizzes, discussions, worksheets, journals, grammar exercises, reading/summarizing articles, listening and taking notes, mini-presentations) • Purpose: Assist in the learning process: assess progress, provide feedback on learning, and inform instruction	• Measuring or summarizing what students have learned and how well students have accomplished objectives; take place at the end of a unit or course (e.g., tests, papers, projects) • Purpose: Determine whether students reached benchmarks and how successful students are in meeting SLOs
Outcome-based assessment	

Table 1. Formative and summative assessments

Objective-Outcome-Assessment Mapping Course: _____			
Objective #	**Corresponding SLO #**	**Summative Assessment**	**Ways of Formative Assessment**

Table 2. A tool for curriculum-assessment mapping

the fact, when the test has already been given and used" (Bachman 2001, 110). The guide is structured as follows. First, it provides specific steps and measures to be considered when planning for courses in terms of assessment (e.g., determining how each course objective will be assessed and under which grading category and percentage during the course planning phase). Second, it outlines important aspects of validity and reliability and provides a checklist with questions and measures to be consulted during the assessment cycle. Finally, it offers a multistep process containing clear procedures and tools to guide the creation, administration, evaluation, and assessment of summative assessments (e.g., final exams).

ASSESSMENT FRAMEWORK

Course Planning Roadmap

The following five steps need to be considered during the course planning phase:

1. Read through the course objectives and SLOs, and brainstorm for means to collect evidence of student learning using formative and summative assessments (see Table 1).

2. Determine how each course objective/SLO will be assessed. There are many tools you may adapt to assist with this process. Fink's (2003, 23) guide to course design offers a backward design template that could be used for this purpose. Teachers can also create a grid to map course objectives to SLOs and to formative and summative assessments, as presented in Table 2. This tool is effective in taking a deeper look into a course curriculum and how it is assessed.

3. After determining means of assessment, identify the grading category that each means of assessment will go under (i.e., the grading criteria subsection of your course syllabus), along with the percentage (if any) it will carry. Remember that formative assessments are low stakes and should have no or a minimal grade point.

4. List major assessments (summative assessments) and plug them in the course calendar so that students are aware of major-assessment events and due dates.

5. Account for and ensure adherence to the fundamental principles of assessments when you plan for an assessment event. Tables 3 and 4 show guiding questions and measures to maximize the validity and reliability of assessment instruments and procedures.

Steps and Guidelines for Summative (Final) Assessments

The following are specific measures to guide the creation of a summative assessment, the administration of the assessment, the grading of the assessment, and the analysis of assessment data.

1. Creating the assessment

Take the following steps *prior to creating the assessment*:

- Determine what skills and knowledge need to be assessed, based on SLOs and instruction received.

- Examine samples of former assessments for a given course, if available.

Do the following *when creating your final assessment*:

- Write clear test instructions and unambiguous assessment tasks for the test blueprint. Decide what acceptable responses should look like and write directions that clearly define those expectations. If there are multiple sections of the same course, collaborate with the other instructors to develop the testing instrument to be used in multiple sections; see Bachman (2001) and Bachman and Palmer (1996) for the components of a test blueprint.

- Specify the grading method (e.g., rubric, key, norm then grade individually, graded collaboratively).

- Mark the SLOs covered in the exam and match assessment questions to SLO numbers as they appear in the course syllabus.

- Create a practice final exam to familiarize students with the exam format, content, and response requirements; discuss acceptable responses with students.

- Walk students through the rubric or grading criteria prior to administering the final.

Take these steps *after creating your final assessment*:

- Find an exam reviewer (ideally another instructor who has experience teaching the course) to do the following:
    - o Take the test.
    - o Check the test against the course SLOs and appropriate test format.
    - o Review the scoring criteria.
    - o Provide written feedback and suggestions on any unclear instructions and ambiguous questions.
    - o Discuss the exam with the preparer and make changes, if needed.

2. Administering the assessment

Make sure to do the following:

- Arrive at least ten minutes prior to the exam start time.

- Check desk arrangements before the exam begins.

- Give students time to read through the directions and questions.

- Write the exam start and end times on the whiteboard.

- Announce the use of a dictionary, notes, or other materials if allowed.

Guiding Questions	Measures
Curricular (Content) Validity ☐ Are the assessment tasks aligned to the curriculum?	✓ Align targeted SLOs to the assessment tasks.
Instructional Validity ☐ Is the content and format of the assessment relative to the instruction received?	✓ Assess what has been actually taught.
Face Validity ☐ Does the test "look" valid to students and instructors by simple inspection?	✓ Create tests valid and credible for their purposes (e.g., a speaking test should be composed of "authentic" speaking tasks).
Construct Validity ☐ Does the test measure the skill/ability it intends to measure? ☐ Is scoring of responses related directly to the language ability (construct) being tested?	✓ Clearly define the abilities (constructs) that an assessment aims to measure and score accordingly (e.g., a listening test should measure listening ability; if spelling is accounted for in the scoring process, that should be made explicit in the test instructions, and spelling should be explicitly taught and practiced prior to the exam).

Table 3. A checklist and measures for assessment validity

APPLICATIONS FOR ONLINE ASSESSMENT

The assessment measures in this article can be extended to distance learning and are equally important to ensure assessment reliability and validity in such an environment. The measures related to the creation and grading of a summative assessment and to the analysis of the assessment data are easily transferable to online settings; however, some adaptations must be made to procedures concerning the administration of summative assessments online.

Teachers can use a platform such as Zoom or Microsoft Teams for summative assessments that require proctoring. Students' cameras need to be on at all times to ensure—as much as possible—that they are not receiving help. Another measure that can be undertaken to prevent students from browsing other content while completing the exam is to have them use their phones to join the virtual exam meeting, then place their phones in a way that allows you to see the students and their computer screens (not too close so that students' answers are not visible to everyone).

Creating a practice final exam using the same final-exam format and medium (e.g., an exam built in a learning management system such as Blackboard or simply in a Word document) and the same proctoring procedures is valuable not only to familiarize students with the process but also to help you tackle any issues that might arise. Conducting a practice exam could ease the online administration of the final exam and ensure a reliable as well as smooth virtual testing experience for students and teachers alike.

- Clarify your cell phone policy before the exam begins.

3. Grading the assessment

Consider doing the following *before grading the assessment*:

- Try to keep grading anonymous. It may be helpful for course instructors to either grade together or have each teacher grade a different section's assessments to increase objectivity.

- If there are multiple sections of the same class, hold a norming session to collaboratively grade at least two tests from each section, using a rubric or pre-set scoring criteria.

- Discuss acceptable responses and decide on any considerations, such as partial credit or dropping an ambiguous question.

Take these steps *after grading the assessment*:

- Sort scores from highest to lowest.

- Conduct item analysis for applicable summative assessments for the group being tested. Item analysis is a process of reviewing individual test items for indices of facility (percentage of students getting each item correct) and discrimination (comparison of good and poor examinee performance on test items); this process shows what students have learned, helps identify corresponding SLOs, and pinpoints problematic questions for revision or replacement. Bailey (1998) provides clear examples of how to conduct an item analysis, and a sample online test-item analysis calculator can be accessed at https://lles.pasco.k12.fl.us/?page_id=7119.

- Use a test analysis and SLO template to sort the results by learning outcomes and record incorrect answers to each question for each student.

Guiding Questions	Measures
Test-Related Reliability ☐ Are the instructions clear, explicit, and level-appropriate? ☐ Are the assessment tasks unambiguous?	✓ After creating an assessment, set aside designated time for a review. ✓ Have another instructor review the assessment instructions and tasks. ✓ After handing back the test papers/rubrics, conduct a post-assessment discussion to get students' feedback on the test and use this to inform future test design.
Student-Related Reliability ☐ Is the length of the assessment appropriate to avoid students getting tired, and can the assessment be completed within the allotted time? ☐ Are students familiar with the assessment format, content, and response requirements? ☐ Have the grading criteria been communicated to students prior to the test? ☐ Are there any observed physical or psychological factors (e.g., fatigue, temporary illness, anxiety) that could have resulted in an "untrue" score of a student's competency/proficiency level?	✓ Design assessment tasks that students are familiar with and appropriate for the target population. ✓ Provide practice with the testing format. Explain response-requirement expectations. ✓ Share and go over the rubric or scoring criteria with students prior to administering a test. ✓ Consider having a clear "Late Work Policy" in the course syllabus that accounts for absences due to sickness and other circumstances (e.g., allowing a makeup and/or dropping the lowest score). Provide reasonable accommodations for students with documented disabilities (consult your institutional policy regarding reasonable accommodations).
Test Administration-Related Reliability ☐ Are the conditions in which the assessment is administered proper (e.g., photocopying, equipment, noise level, room temperature)?	✓ Ensure clearly typed and cleanly photocopied exams. ✓ Test the equipment (e.g., computers, audio, projector) before the assessment. ✓ Request another classroom to administer a test in if the noise level is high or room temperature is unbearable.
Interrater-Related Reliability ☐ Have scoring criteria been set? ☐ Has grading on selected criteria and benchmarks been done jointly or agreed upon with other teachers? ☐ Are the scores consistent across multiple evaluators? ☐ Are the scoring procedures consistent (e.g., giving partial credit to all students)? ☐ Have measures been undertaken to ensure a scoring process free of bias, subjectivity, and human error?	✓ Set clear scoring criteria; rubrics for speaking and writing assessments; and a scoring guide and answer keys for listening, reading, and grammar assessments. ✓ Have more than one evaluator and/or hold a norming session. ✓ If scores are inconsistent across multiple evaluators, have them discuss the reasoning behind a specific score, come to a consensus, and adjust the grading criteria if necessary. ✓ Ensure consistent scoring procedures. Read through all tests before scoring. Determine acceptable responses before scoring starts. ✓ Grade anonymously and review scores before making them available to students. ✓ Compile assessment (numerical) data through analyzing individual test scores to provide feedback to individual students, inform instruction, and improve the curriculum. Consult an online test-item analysis calculator for this process.

Table 4. A checklist and measures for assessment reliability

The assessment measures in this article can be extended to distance learning and are equally important to ensure assessment reliability and validity in such an environment.

4. Analyzing the assessment data

Assess your assessment by determining pass/fail rates and identifying problematic questions. Teachers are highly encouraged to close the assessment loop by analyzing assessment data and using the results to improve student learning (Banta and Blaich 2010; Schoepp and Tezcan-Unal 2017). NOVA (2018) contains information and resources about the "continuous nature of assessing student learning outcomes," including "closing the loop."

To illustrate, after conducting the test-item analysis mentioned above, teachers can summarize student performance on the final assessment and share the data with curriculum/program administrators. This perhaps could be realized by making the closing of the assessment loop a required procedure at the program level for continuous improvement. To facilitate this, course instructors can create a final-exam evaluation form to (1) indicate the percentage or total raw number of students meeting the SLOs covered in the final assessment; (2) identify particular questions and corresponding SLOs that more than half of the class answered incorrectly; and (3) offer a brief rationale for why the course instructor thinks those questions were problematic for students and what might be done in the future. As described by Suskie (2018), sometimes the reason for poor performance on a test is clear (e.g., unclear prompt, ambiguous wording), but sometimes it is not as obvious (e.g., flawed curriculum design, ineffective teaching methods).

Documenting and sharing assessment results transparently informs decisions regarding adding or adjusting assignments or instructional materials, revising assessment methods, modifying specific course objectives and outcomes, revising course sequencing, and looking into opportunities for faculty training and professional development.

FINAL REMARKS

It is important to note that assessment might be handled differently from one program to another. Some programs have a position designed to prepare and develop summative-assessment instruments for their faculty, while in other programs, course instructors design their own assessments. Regardless, the proposed procedures in this article can benefit assessment coordinators, individual teachers, and programs at large. ESL/EFL programs seeking accreditation or standardizing their assessment procedures can adapt some of the measures outlined in this article to ensure that testing is reliable and valid.

Even taking small steps—such as seeking an exam reviewer to provide input on an exam (peer-reviewed final exams), grading collaboratively and anonymously, linking test items to SLOs, and analyzing final-assessment data—ensures that assessment instruments and procedures are fair, accurate, and consistent in assessing learning. Formalizing these small yet powerful steps and establishing a systematized set of procedures to guide instructors across the assessment cycle optimize assessment effectiveness.

REFERENCES

Bachman, L. 2001. Designing and developing useful language tests. In *Experimenting with uncertainty: Essays in honour of Alan Davies*, ed. C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley,

T. McNamara, and K. O'Loughlin, 109–116. Cambridge: Cambridge University Press.
Bachman, L., and B. Damböck. 2018. *Language assessment for classroom teachers*. Oxford: Oxford University Press.
Bachman, L. F., and A. S. Palmer. 1996. *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.
Bailey, K. M. 1998. *Learning about language assessment: Dilemmas, decisions, and directions*. Boston: Heinle & Heinle.
Banta, T. W., and C. Blaich. 2010. Closing the assessment loop. *Change: The Magazine of Higher Learning* 43 (1): 22–27.
Birenbaum, M. 2007. Evaluating the assessment: Sources of evidence for quality assurance. *Studies in Educational Evaluation* 33 (1): 29–49.
CEA. 2019. "Standards." Commission on English Language Program Accreditation. https://cea-accredit.org/about-cea/standards
Fink, L. D. 2003. A self-directed guide to designing courses for significant learning. Instructional Development Program: University of Oklahoma, 1–35. https://www.deefinkandassociates.com/GuidetoCourseDesignAug05.pdf
NOVA. 2018. "Assessment Loop Resources." Northern Virginia Community College. https://www.nvcc.edu/assessment/loop/index.html
Priyanti, N. 2017. Test usefulness of IELTS writing test tasks. *Apples – Journal of Applied Language Studies* 11 (4): 1–9.
Schmidgall, J. E., E. P. Getman, and J. Zu. 2018. Screener tests need validation too: Weighing an argument for test use against practical concerns. *Language Testing* 35 (4): 583–607.
Schoepp, K., and B. Tezcan-Unal. 2017. Examining the effectiveness of a learning outcomes assessment program: A four frames perspective. *Innovative Higher Education* 42 (4): 305–319.
Suskie, L. 2018. *Assessing student learning: A common sense guide*. 3rd ed. San Francisco: Jossey-Bass.
Thu, D. M. 2019. A review on validating language tests. *VNU Journal of Foreign Studies* 35 (1): 143–154.

Eman Elturki has taught ESL/EFL for over 15 years and has held administrative positions related to curriculum, assessment, and accreditation. In addition to teaching in pathway programs at Washington State University, Elturki served as Senior Associate Director of Academic Programs.