Classroom Assessment: A View from a Secondary ESL Teacher

Irina Bleckhman, Reynolds High School

When I began teaching ESL in Oregon's secondary public schools 15 years ago, my main formal classroom assessments were weekly quizzes and end-of-unit tests which I designed using a combination of multiple-choice, cloze, matching, and open-ended items. What I wanted to know was how well my students knew the content of their ESL curriculum, and I believed the information from quizzes and tests was sufficient for me to draw some conclusions about their overall language proficiency. Besides the students themselves and some of their parents, few people were interested in these conclusions.

Today, due to the changes brought to K-12 ESOL/Bilingual education by the No Child Left Behind Act and the subsequent state-level mandates, my students' English proficiency is of high interest to many more people. New policies are centered exclusively around the notion of English proficiency. One of them is the standardized Oregon English Language Proficiency Assessment (ELPA). This test is given to all English language learners in K-12 public schools each year, and the results are closely monitored at the school, district, state, and even federal levels. The other policy is reflected in the goals set forth by the Oregon Department of Education and my school district. To meet these goals, 65 percent of my ESL students must advance to the next level of English proficiency this school year, and within three years 95 percent need to advance at least one level annually.

The heightened urgency to enhance the rate of language acquisition brought on by recent education reforms has had a significant impact on my approach to classroom assessment. Today I need to have ongoing assessment of my students that not only allows me to make conclusions about their mastery of the narrow instructional goals of each individual lesson, but that also gives me accurate, comprehensive, and current information about students' overall proficiency in English. Classroom performance assessments have become my preferred form of evaluation. This article will begin with a definition of language proficiency. Next will be a description of some institutional and technical challenges to classroom performance assessment, then examples from my classroom.

Language Proficiency and Assessment Instruments

The description of language competence reflected in various institutional aspects of my program is close to that posted on the website of the Oregon Department of Education and which was adapted from Bachman (1990). (See Figure 1 below.) Figure 1 presents language competence as a dynamic combination of various other competencies that interact and contribute to one's ability to communicate using language in specific social contexts. It views language proficiency as socially situated rather than something that belongs entirely to an individual.

The description is institutionalized in the English Language Proficiency Standards (Oregon Department of Education, 2009). The Standards have the same ability-based and task-based orientation that Bachman's definition implies.

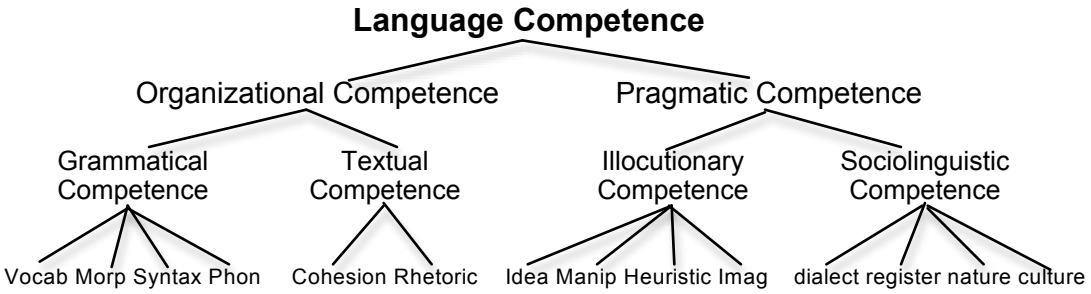

Figure 1. Definition of language proficiency from Backman (1990) and used by the Oregon Department of Education (2004).

Having been developed with the purpose of describing the language needs of students at various levels and prescribing language curriculum objectives, the Standards represent a hierarchical approach to the construct of proficiency and have a strong focus on academic tasks. The following is an example of one of the standards related to vocabulary knowledge:

> EL.HS.RE.08 Understand, learn, and use new vocabulary that is introduced and taught directly through informational text, literary text, and instruction across the subject areas.
>
> Advanced
> Interprets words appropriately that sometimes have multiple meanings and applies this knowledge consistently to literature and texts in content areas.
>
> Early Advanced
> Recognizes that words sometimes have multiple meanings and applies this knowledge to understanding written texts.
>
> Intermediate
> Recognizes that words sometimes have multiple meanings and applies this knowledge to understanding written texts.
>
> Early Intermediate
> Recognizes that words sometimes have multiple meanings.
>
> Beginning
> Produces simple vocabulary (single words or short phrases) to communicate basic needs in social and academic settings (e.g., locations, greetings, classroom objects). (p.91)

The instructional methodology adopted by my program reflects the institutional definition of language proficiency as communicative and socially situated language ability. This methodology is known to many Oregon K-12 ESL teachers as Focused Approach to Systematic English Language Development, or FASELD (Dutro, 2008). FASELD is the work of E. L. Achieve, an educational consulting company. This approach carries many features of Focus-on-Form (FonF) instruction (Doughty & Williams, 1998; Spada, 1997; Williams, 2005).

Contrary to the name, the main emphasis of FonF language instruction is on meaning, but it shifts toward form at certain stages of the instructional process. FASELD, as it is used in my program, is probably the most teacher-controlled version of FonF since it is planned (activities are designed with the goal of bringing learners' attention to specific forms), proactive (forms are selected based often on a pre-established profile of the learner's interlanguage), and targeted (communicative tasks are highly focused around forms) (Williams, 2005). The textbook adopted by the curriculum of my program, *Top Notch* (Saslow & Ascher, 2000), builds instruction around specific communicative competencies that span a wide range of lexico-grammatical content and sociolinguistic contexts.

Institutional Challenges

Establishing a single classroom process that serves placement, evaluative, formative, and summative purposes has been a significant

challenge. My goal has been to integrate performance assessments of proficiency organically into teaching and learning while having them also serve the evaluative purpose usually reserved for classroom tests. Hughes (2003) suggested that achievement tests can become reasonable measures of proficiency if they meet several conditions. First, achievement tests should be based on the objectives of the course rather than being closely tied to the detailed content of the course and its materials. Second, course objectives have to be based on the real language needs of the students in relation to their proficiency level. Finally, if multiple assessments are used at the end of each instructional unit instead of one overall assessment at the end of the year, their success depends in large part on how well the short-term objectives of each instructional unit represent the overall goal of a specific proficiency level.

The structure of my program makes it fairly easy to address the second and third requirements. Course objectives are based on a pre-established profile of students' interlanguage. This profile is reflected in proficiency-level descriptors widely accepted by the school district and heavily utilized in program planning (Oregon Department of Education, 1994: 5-6). The short-term goals of each instructional unit, supported by curriculum and methodology, represent a systematic organization of communicative competencies that span a variety of social contexts. There is a wide consensus among practitioners in my program and experts in the field that these goals lead to the next proficiency level or full proficiency in English, thus fulfilling Hughes' (2003) third requirement that achievement tests serve as measures of proficiency.

The first requirement, however, presents a real challenge in language performance assessments. It calls for achievement tests to be tied to course objectives more than to specific content. In other words, achievement tests that serve as measures of proficiency should assess how well students are able to perform a communicative task with any linguistic content, not just the specific vocabulary, grammar, and sociocultural information they have received in the classroom. However, the connection to course content provides information about the relationship between teaching and learning in the classroom, so it has to play a certain role in assessment.

Assessing too much specific course content does not provide valid information about a student's ability to perform a communicative task in all of its complexity. Assessing too little allows quite a few students to complete the task by relying on pre-existing competencies. As examples in Appendix A demonstrate, the performance prompts can be flexible while limiting vocabulary choices to the targeted vocabulary of the lesson and specifying which grammatical forms students have to use.

Another significant challenge to performance assessment comes from environmental factors: class size, the limited amount of time available for assessment, and the lack of access to technology. Writing becomes the most efficient form of assessment, but it limits linguistic performance to only one domain. Additionally, writing is a complex skill comprised of linguistic and many non-linguistic aspects (Kroll, 1990; Weigle, 2002). Some of my secondary students struggle with the non-linguistic demands of writing too much to adequately demonstrate their linguistic competencies.

To address this issue, for each performance assessment I select 3-5 students to provide oral responses instead of written ones. By the end of the year, each student will have performed approximately 30 percent of the assessment tasks orally. These oral performances are usually spaced out throughout the year, which allows me to assess growth in oral proficiency in addition to writing proficiency. Several students who have especially low writing skills due to interrupted formal schooling or learning disabilities are assessed orally most of the time. Finally, a writing scoring rubric can exclude non-linguistic aspects of the performance (see Appendix B).

Many other challenges in integrating language performance assessments into teaching and learning have arisen. These include determining the optimal place of such assessments in the instructional cycle, finding a fair way to factor them into class grades, getting students motivated to do their best, finding the most effective record-keeping system, and figuring out the most meaningful and effective ways to analyze the data.

Technical Challenges

Designing assessment instruments and establishing well-integrated assessment processes require continual examination of their quality. The psychometric tradition has relied heavily on the notions of validity and reliability to discuss the quality of large-scale language assessments. However, there is some agreement in the scholarly community that these notions may not be applicable to classroom performance assessment (Leung, 2005). Classroom assessments have certain inherent qualities (e.g., their variability, context-centeredness, and innate authenticity) that make them methodologically different from large-scale language tests and unlikely subjects of analyses conducted from the traditional psychometric point of view. Classroom assessment also represents an entirely different epistemological approach to language testing (Huerta-Macias, 1995; Gipps, 1994). Additionally, it is argued that classroom assessments do not need to be held to the same standard of quality as standardized measures since the higher authenticity and contextuality in classroom assessments already make them more credible, and decisions that are made based on the results are generally low-stakes.

I believe that classroom assessments are indeed epistemologically and methodologically different from large-scale standardized measures and should be conceptualized differently. However, due to the institutional factors described earlier, the decisions we in Oregon make based on classroom assessments are not exactly low-stakes. These assessments are also not necessarily inherently authentic since most of the tasks have contexts that are largely simulated or imagined. This raises my concern about the quality of the instruments I use.

Since there are real difficulties in applying the psychometric perspective to classroom performance assessments due to their lack of standardization, alternative terms such as trustworthiness (Huerta-Macias, 1995), credibility, and auditability have been offered for conceptualizing the validity and reliability of these assessments. I will discuss quality issues using the traditional notions of validity and reliability, keeping in mind that the processes of establishing and maintaining these characteristics in classroom assessments are fundamentally different from those in standardized testing. I will also discuss practical measures that I take to enhance the quality of my assessment instruments.

Content validity

My students' performances on these classroom assessments need to be indicative of their level of proficiency as defined by the construct described earlier. In other words, I want these assessments to have high content validity. One way to establish the content validity of these assessments is to design representative tasks or tasks that are very likely to lead to performances in which my students demonstrate their true language abilities (Hughes, 2003). But how do I know if my tasks are representative? Chief among the practical measures to take to address this validity concern is to consult with colleagues who are experts in the field and familiar with the institutional aspects of the construct of proficiency.

I can also seek evidence about how my students interpret the content of the assessment tasks by giving each of my instruments a trial run. If I find evidence that my students interpret the content of the assessment in ways that were not intended, I can check assessment items and scoring rubrics for clarity and the level of restrictiveness. During the trial run, I can also look

for evidence of my instrument's ability to clearly discriminate between stronger and weaker language learners. For this reason, I might administer my assessments to a few native speakers of the same age group. If I find evidence that my assessments fail to clearly discriminate between more proficient language learners, especially native speakers, and less proficient ones, I can safely assume that the task needs to be revised.

Another concern related to content validity is the generalizability of student performance. The conclusions about proficiency made from classroom assessments should be based on an accurate representation of their ability to perform the tasks in a variety of contexts, including contexts outside of school. This is difficult since the level of performance varies with the task, often significantly (Linn, Baker, & Dunbar, 1991). One obvious way to increase generalizability is to increase the number of performances assessed. Another way is to make sure that the tasks represent "systemically critically dimensions" of the construct (Linn, et al., 1991: 19). By allowing performance-based assessments to take the place of achievement tests at the end of each instructional unit, I can assess close to 20 performances by each of my students in the course of the year. Since these samples are closely tied to the curriculum, which already represents systemically critical dimensions of the construct, I can be reassured that their generalizability is fairly high.

Not unlike many other language teachers, I take other measures without ever formalizing them in order to ensure the validity of my assessments. These measures are representative of the epistemological differences between large-scale testing and classroom assessments mentioned earlier. For example, instead of conducting a formal study of concurrent validity, I compare what a student was able to do on her performance assessment with multiple pieces of other evidence I have from her other in-class or, perhaps, extra-curricular activities. Instead of formally comparing language samples from the performance assessments in simulated contexts with language samples obtained in comparable but authentic contexts and making conclusions about generalizability, I compare my students' performance to the abilities they demonstrate communicating with me or with each other outside of class. While this is a much more holistic approach to validating assessments, it responds to some of the same concerns as a research study of validity would.

Reliability

Cohen (1994) pointed out that "three different types of factors contribute to the reliability of language assessments: test factors, situation factors, and individual factors" (p.36). Many of the test factors can be addressed along with concerns associated with validity through wide sampling, examining student perceptions of content, ensuring clarity of items, specifying content, and collaborating with other experts. If there are problems with reliability, they are visible through large discrepancies between individual performances by the same student, large discrepancies between performance on instructional tasks and assessment tasks, and discrepancies between groups of students identified based on factors unrelated to English proficiency. With multiple sources of information available, I often find it unnecessary (as well as impractical) to do formal reliability studies. However, I have occasionally used a variation of a split-half method, asking my students to respond to two prompts that tested the same communicative competence. I have also used a score-rescore method to ensure my own intra-rater reliability.

Situation factors can be addressed through making testing situations uniform for all performances (for example, all of my assessments must be completed individually and independently in class with ample time for students to prepare their responses). I work to control individual factors by using my knowledge about each student to determine the time for administering assessments, the length of assessments, and my students' motivation for a particular task.

Fairness

Addressing validity and reliability also helps with fairness. But there is more. Secondary ESL learners are an extremely diverse group. In addition to the cultural diversity expected among ESL students, secondary ESL learners vary greatly in their academic experiences. Some have had adequate schooling; others, limited or interrupted schooling (Freeman, Freeman, & Mercuri, 2002). Some are recent arrivals, and some were born in the United States. Some have had all of their schooling in English, while others have been schooled mostly in their native languages. These students show similar results on standardized assessments and are put in the same proficiency groups for instruction, but their learning needs are quite diverse.

No classroom assessment is completely devoid of bias. Each of my assessments requires a certain type of background cultural knowledge and a certain level of academic skill, two areas in which my students vary widely. However, I can estimate with some accuracy when bias would compromise fairness for each student. For example, if I know that a student is a recent arrival from a remote village in his native country and has never driven a car, it would be unfair to ask that student to describe the process of obtaining a driver's license in order to assess his ability to describe a sequence of actions. If I know that a student's academic skills are significantly below grade level, it would be unfair to ask her to describe a complex social phenomenon in order to assess her ability to explain cause and effect in detail.

What I can do in my assessments is to control bias through my knowledge of my students and to compensate for it through instruction. While standardized assessments aim to exclude students' individual experiences from the test, I often aim to include them. If I know that most of my students are familiar with a certain cultural context (knowledge I obtain through classroom interactions), I include that context in my assessment task. If I know that their performance on a classroom assessment will depend on a particular skill, I teach that skill. Thus the contextual nature of classroom assessments does not preclude me from addressing concerns about fairness. On the contrary, it gives me tools that are already embedded in teaching and learning processes to understand and mitigate cultural and experiential biases. The example in Figure 2 represents the type of constrained constructed response that I most often use in my performance assessments. The task in the example approximates a real-life situation of giving health-related advice. Even though grammatical competence is singled out and assessed as a separate trait, the overall ability to perform the task is also assessed. Appendix A offers some of the performance assessments for the first semester of an intermediate-level course in order to show the scope of sampling.

Conclusion

The institutional factors I have described set concrete parameters that apply to assessment of proficiency in the classroom. Language proficiency is viewed as the socially situated ability to perform a wide variety of communicative tasks, and this ability is comprised of multiple competencies that work in an integrated fashion.

Short-term Curriculum Objective: Students will be able to give health advice using their knowledge of vocabulary for common illnesses and pharmacy produts and their knowledge of modals of necessity.

Student prompt: David sprained his wrist. Karim has a headache. Victoria has the flu. Give each person advice. Use the names of pharmacy products and modal verbs in each response.

Advice for Daniel:

_____

_____

Advice for Karim:

_____

_____

Advice for Victoria:

_____

_____

Figure 2. Performance assessment for an intermediate-level secondary ESL course

As a result, the most direct and therefore the most valid assessment of language proficiency requires an integrative assessment of linguistic performance (Hughes, 2003). Another requirement is that a comprehensive assessment of language proficiency take place across an array of various constructed or simulated social situations. What this means for the technical aspect of assessment design is that classroom assessments need to allow multiple observations of students' linguistic performances while they are engaged in varied tasks and contexts.

Classroom assessments have new significance in secondary ESL programs. Their purpose, shaped by institutional factors, is often not only to measure the degree of success with which students have mastered narrow learning objectives of each lesson, but also to measure language proficiency. Despite, and in some cases due to, methodological and epistemological differences between standardized tests and classroom assessments, classroom performance assessments are able to serve as quality measures of proficiency if they properly reflect proficiency-oriented objectives and adequately address concerns about validity, reliability, and fairness through the instructional process.

References

Bachman, L. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press, 1990.

Cohen, A. (1994), *Assessing language ability in the classroom* (2nd ed.). Boston: Heinle & Heinle.

Doughty, C., & Williams, J. (Eds.) (1998). *Focus on form in classroom second language acquisition*. Cambridge: Cambridge University Press.

Dutro, S. (2009). *Systematic English language development: A handbook for secondary teachers*. Santa Cruz, CA: ToucanEd.

Ellis, N. (2001). Memory for language. In P. Robinson (Ed.), *Cognition and second language instruction* (pp.33-68). Cambridge: Cambridge University Press.

Freeman, Y., Freeman, D., & Mercuri, S. (2002). *Closing the achievement gap: How to reach limited-formal schooling and long-term English learners*. Westport, CT: Heinemann.

Gipps, C. (1994). *Beyond testing: Towards a theory of educational measurement.* London: Falmer Press.

Huerta-Macias, A. (1995). Alternative assessment: Responses to commonly asked questions. *TESOL Journal, 5,* 8-11.

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge, UK: Camridge University Press.

Kroll, B. (Ed.) (1990). *Second language writing: Research insights for the classroom*. Cambridge, UK: Cambridge University Press.

Leung, C. (2005). Classroom teacher assessment of second language development: Construct as practice. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 869-888). Mahwah, N.J: Lawrence Erlbaum Associates.

Linn, R., Baker E., & Dunbar, S. (1991). Complex performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20* (8), 15-21.

McNamara, T.F. (2005). Introduction: The social turn in language assessment. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 775-778). Mahwah, N.J: Lawrence Erlbaum Associates.

Oregon Department of Education. (2004). *English language proficiency standards.* Retrieved from http://www.ode.state.or.us/teachlearn/standards/elp/files/competencechart.pdf.

Oregon Department of Education. (2009*). Standards by design: Seventh grade, eighth grade.* Retrieved from http://www.ode.state.or.us

Saslow, J., & Ascher, A. (2006). *Top Notch: English for today's world*. *Student Book 2*. White Plains, NY: Pearson Education.

Spada, N. (1997) Form-focused instruction and second language acquisition: A review of classroom and laboratory research. *Language Teaching Abstracts*, *30*, 73-87.

Weigle, S. C. (2002). *Assessing writing*. Cambridge, UK: Cambridge University Press.
Williams, J. (2005). Form-focused instruction. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 671-691). Mahwah, N.J: Lawrence Erlbaum Associates.

Irina Blekhman has taught ESOL in secondary public schools for the past fifteen years. She was also an adjunct faculty member at Lewis and Clark Graduate School of Education and Counseling. Her interests include language assessment, education of long-term English language learners, and critical pedagogy in language teaching.

Appendix A

Sample performance assessment tasks for the first semester of an intermediate-level secondary ESL course

Unit 1 Lesson 1	What interesting things have you done? What interesting places have you been to? What would you like to do that you haven't done yet? Respond using present perfect in at least 4 of your sentences.
Unit 2 Lesson 2	Pretend that you are a tourist on your first visit to Portland, Oregon. Before you came to Portland, you made a list of all the things you wanted to do on your trip (list provided). Today is the fourth day of your trip. You have already done some things on your list, but there are some things you have not had time to do yet. You have put a check mark (√) next to the things you have done. Write a post card to your family describing what you have and have not done. Use present perfect and the words *already* and *yet* when you describe your activities.
Unit 3 Lesson 2	Write a paragraph about any topic. Use at least 5 phrasal verbs from the list in your paragraph. Use at least 2 phrasal verbs with pronoun objects (for example, *pick them up, turn it on*). Phrasal verbs: turn on, put up, get away with, turn off, put off, get along with, turn down, put down, pick on, turn up, put away, pick up, put on, drop off , turn out, pick out, put up with, take on, get up, take in, take away, get off
Unit 4 Lesson 1	What personal care products do you use? How often and where do you buy them? Why do you like them? Write a paragraph using as many names of personal care products as you can and the words *many, much, a lot of, some, a/an,* and *any* before nouns.
Unit 4 Lesson 3	Daniel sprained his wrist. Karim has a headache. Victoria has the flu. Give each person some advice. Use the names of pharmacy products and modal verbs in each response.

Appendix B

Scoring Rubric

	0: No mastery	1: Partial mastery	2: Mastery
How effectively does the student complete the communicative task?	Does not attempt to perform the task, provides a response that is off-topic, or fails to communicate in a manner that can be understood.	Addresses only a portion of the prompt or communicates in a way that can be only partially understood.	Addresses most aspects of the prompt in a way that can be easily understood.
Topic-related vocabulary	Does not use topic-related vocabulary.	Uses some topic-related vocabulary correctly, but the knowledge of vocabulary is insufficient (does not use the right item when there is a clear need to do so) or incomplete (some vocabulary is misused).	Addresses most aspects of the prompt in a manner that can be easily understood.
Grammar	Does not use targeted grammatical forms or uses all targeted grammatical forms incorrectly.	Some errors in targeted grammatical forms; lack of consistency.	All targeted grammatical forms are used correctly.