Second Language Assessment for Classroom Teachers

Thu H. Tran

Missouri University of Science and Technology

Rolla, Missouri, USA

Abstract

The vast majority of second language teachers feels confident about their instructional performance and does not usually have much difficulty with their teaching thanks to their professional training and accumulated classroom experience. Nonetheless, many second language teachers may not have received sufficient training in test development to design sound classroom assessment tools. The overarching aim of this paper is to discuss some fundamental issues in second language assessment to provide classroom practitioners with knowledge to improve their test development skills. The paper begins with the history, importance, and necessity of language assessment. It then reviews common terms used in second language assessment and identifies categories and types of assessments. Next, it examines major principles for second language assessment including validity, reliability, practicality, equivalency, authenticity, and washback. The paper also discusses an array of options in language assessment which can generally be classified into three groups: selected-response, constructed-response, and personal response. Finally, the paper argues for a balanced approach to second language assessment which should be utilized to yield more reliable language assessment results, and such an approach may reduce students' stress and anxiety. (This paper contains one table.)

The History of Language Assessment

It may be difficult to find out when language assessment was first employed. The Old Testament, however, described an excellent example of one of the earliest written documents related to some kind of language assessment, as noted by Hopkins, Stanley, and Hopkins (1990).

> The Gileadites took the fords of the Jordan toward Ephraim. When any of the fleeing Ephraimites said, "Let me pass," the men of Gilead would say to him, "Are you an Ephraimite?" If he answered, "No!" they would ask him to say "shibboleth." If he said "sibboleth," not being able to give the proper pronunciation, they would seize him and kill him at the fords of Jordan. [Judges 12:5-6]

The description above in the Bible may be considered one of the most extreme forms of high-stakes tests in the history of language testing. The pronunciation test depicted is fairly clear. Being able to pronounce the "h" sound in the word "shibboleth" meant that one was Gileadite, and his life would be saved. Otherwise, he would be killed. These days language assessment is of importance as it is employed in a variety of contexts for various purposes and with different measuring devices and methods to help determine the level of language proficiency of learners. Decisions based on language proficiency assessment may at times have important implications on students' academic and professional lives.

The Importance and Necessity of Assessment

In the field of education, "some form of assessment is inevitable; it is inherent in the teaching – learning process" (Hopkins, Stanley, & Hopkins, 1990, p.194). In a similar vein, Stoynoff and Chapelle (2005) stated that "teachers are involved in many forms of assessment and testing through their daily teaching and use of test scores" (p. 1), but they also noted that many teachers find principles of assessment an aspect that is difficult to update and apply efficiently.

These authors also indicated that although teachers can construct tests and test specialists can teach classes, the roles and daily activities of the two groups are different. Although the roles of teachers and testers are clearly differentiated, it is almost impossible to assess students' academic progress without teachers. In effect, Hopkins, Stanley, and Hopkins (ibid) noted that classroom teachers play "a constant evaluative role" (p. 194) because they have to attempt to decide on students' degree of scholastic achievement and growth.

In reality, teachers working in institutions where there are no standardized or institutionally prepared tests have to construct their own tests for their classes, and when it is the case, the tests constructed by teachers may not be as well designed as those written by professional testers. The author of this paper conducted a small scale survey of the top 10 programs that provide Master's Degree in Teaching English to Speakers of Other Languages (MA TESOL) in the United States by Google search and it was found that only four of the programs provided a course on language assessment or evaluation as a required course and one offered a course on assessing English language learners as an elective course. Half of the MA TESOL programs surveyed did not provide any courses related to language assessment or evaluation. Coombe, Folse, and Hubley (2007) might have been correct in observing that assessment is foreign territory for many teachers. It is, therefore, of great importance and necessity to provide second language classroom practitioners with some fundamental principles and methods of testing, as not all TESOL classroom practitioners are formally trained in second language assessment. Even when teachers are trained in language assessment, keeping abreast of current developments in second language assessment can be a challenge. The primary purpose of this paper is to discuss the principles of assessment and major language assessment types and options to enable classroom language teachers to have a better understanding of constructing effective classroom tests.

Assessment Terminology

Common terms teachers are familiar with may be measurement, test, evaluation, and assessment. The aforementioned terms may informally be used interchangeably to refer to the practice of determining learners' language proficiency in a variety of contexts. However, Bachman (1990) defined measurement in the social sciences as "the process of quantifying the characteristics of persons according to explicit procedures and rules" (p. 18). In education, measurement is "the process of quantifying the observed performance of classroom learners" (Brown & Abeywickrama, 2010, p. 4). Brown and Abeywickrama also mentioned that students' performance can be described both quantitatively and qualitatively, or by assigning numbers such as rankings and letter grades or by providing written descriptions, oral feedback and narrative report.

Drawing from the definition from Carroll (1968), Bachman (ibid) stated that a test is "a measurement instrument designed to elicit a specific sample of an individual's behavior" (p. 20). Similarly, Brown and Abeywickrama (2010) saw tests as a way of measuring a person's ability, knowledge, or performance in a specific domain.

Citing from Weiss (1972), Bachman (ibid) noted that "evaluation can be defined as the systematic gathering of information for the purpose of making decisions" (p.22). Bachman (ibid) also added that one part of evaluation is "the collection of reliable and relevant information" (p. 22). Evaluation involves the interpretation of testing results used to make decisions (Brown & Abeywickrama, 2010). The example provided by Brown and Abeywickrama (ibid) is that "if a student achieves a score of 75 percent (measurement) on a final classroom examination, he or she may be told that the score resulted in a failure (evaluation) to pass the course" (p.5).

Mihai (2010) asserted that assessment is "much more than tests and test scores" (p. 22). Assessment, according to Mihai, is a combination of all kinds of formal and informal judgments and findings occurring inside and outside a classroom. In An Encyclopedic Dictionary of Language Testing, Mousavi (2009, p. 36) defined assessment as "appraising or estimating the level or magnitude of some attribute of a person." Assessment, as Brown and Abeywickrama (2010) added, is an ongoing process including a wide range of techniques such as simply making an oral appraisal of a student's response or jotting down a phrase to comment on a student's essay. Brown and Abeywickrama (ibid) also stated that "a good teacher never ceases to assess students, whether those assessments are incidental or intended" (p. 3).

Categories of Evaluation and Assessments

Evaluation

In discussing about language program evaluation, Richards (2001) presented three types of evaluation: formative, illuminative, and summative evaluation. Formative evaluation, as Richards pointed out, is utilized to find out the aspects of a program that are working well, not working well, and issues that need to be addressed. Some questions related to formative evaluation may involve seeking to find out if enough time has been spent on certain objectives or if the learning materials are well received. For classroom teachers, formative evaluation is an ongoing formal or informal evaluative process in which students are provided with various types of quizzes or tests which serve as a means for student learning. Illuminative evaluation, according to Richards, is employed to find out how different aspects of a program are implemented and this type of evaluation is one way to seek to have "a deeper understanding of the process of teaching and learning that occur in the program, without necessarily seeking to change the course in any way as a result" (ibid, p. 289). According to Passerini and Granger (2000), "illuminative evaluations disclose important factors and issues emerging in a particular learning situation, factors which might have been overlooked by the instructor" (p.13). Examples of illuminative evaluation that Richards provided are finding out the strategies for error-correction teachers employ or the strategies students use to deal with different text types. Illuminative evaluation for classroom teachers can be an instrument designed to assess a specific language point or problem to have a better understanding about students' difficulty in acquiring it, so that appropriate actions can be made. Finally, Richards indicated that summative evaluation, the kind of evaluation most teachers and administrators are familiar with, is concerned with determining the effectiveness, efficiency, and to some extent the acceptability of a language program. Questions related to summative evaluation are if the course achieved its aims, what students learned, and if appropriate teaching methods were used. Summative evaluation can usually be final tests for classroom teachers.

Assessment

For classroom assessment, Mihai (2010) categorized it according to intention, purpose, interpretation, and administration. In regard to intention, an assessment can be informal when it is a spontaneous comment, or it can be formal when it is carried out in a systematic manner. In terms of purpose, an assessment can be formative if it focuses on the process of learning or it can be summative when it is used to measure student learning outcomes at the end of an education cycle. With respect to interpretation, an assessment may be used to compare students' performance with their peers' performance (norm-referenced) or it may be employed to compare students' performance with the course content (criterion-referenced). Mihai (ibid) clarified that "whereas norm-referenced tests evaluate students in terms of their ranking to another, criterion-referenced tests evaluate students in terms of their mastery of course content" (p. 31). The last category of assessment Mihai presented is administration which refers to the way an assessment is administered or delivered; an assessment may be classroom-based (small scale) when it is only used in the classroom or it can be delivered statewide or nationwide (large scale). Assessment, moreover, can be conducted by either speaking or writing. Therefore, one more category of assessment may be added to those provided by Mihai: mode (oral or written). Table 1 provides a summary of types of assessment built upon the one provided by Mihai.

Table 1: The categories and types of assessment

Category of Assessment	**Type of Assessment**
Mode	Oral Written
Intention	Informal Formal
Purpose	Formative Summative
Interpretation	Norm-referenced Criterion-referenced
Administration	Classroom-based Large scale

Principles of Second Language Assessment

Fundamental principles for evaluating and designing second language assessment include validity, reliability, practicality, equivalency, authenticity, and washback.

Validity

A test is considered valid when it reflects the test-takers' ability in a particular area and the test does not measure anything else. Validity is a complex concept in testing, but Brown and Abeywickrama (2010, p. 30) seemed to have well encapsulated the main attributes of validity. They indicated that in order to achieve validity a test should:

- Measure only what it claims to measure,
- Not measure anything else,
- Rely as much as possible on empirical evidence,
- Involve performance that samples the test criterion,
- Offer meaningful and useful information about a test-taker's ability,
- Be supported by a theoretical rationale.

Reliability

A test is considered reliable if it is administered on different occasions and similar results are obtained. Brown and Abeywickrama (2010, p. 27) suggested the following ways to ensure that a test is reliable:

- It is consistent in its conditions across two or more administrations.
- It gives clear directions for scoring or evaluation.
- It has uniform rubrics for scoring or evaluation.
- It lends itself to consistent application of those rubrics by the rater.
- It contains items or tasks that are unambiguous to the test-takers.

Practicality

Practicality refers to the logistical, practical, and administrative issues involved in the process of constructing, administering, and rating an assessment instrument (Brown & Abeywickrama, 2010). Bachman and Palmer (1996, p. 36), on the other hand, defined practicality as "the relationship between the resources that will be required in the design, development, and use of the test and the resources that will be available for these activities." Bachman and Palmer also added that practicality refers to the extent to which the demands of test specifications can be met within the limits of existing resources such as human resources (test writers, raters, or proctors), material resources (space, equipment, or materials), and time.

Equivalency and Authenticity

"An assessment has the property of equivalency if it is directly based on curriculum standards or instructional activities. Specifically, equivalency determines in what ways assessment design is influenced by teaching" (Mihai, 2010, p. 45). Equivalency is somewhat similar to authenticity which is defined as "the degree of correspondence of the characteristics of a given language test task to the features of a target language task" (Bachman & Palmer, 1996, p. 23). Brown and Abeywickrama (2010) provided characteristics of a test that has authenticity as follows:

- It has language that is as natural as possible.
- It contains items that are contextualized rather than isolated.
- It includes topics that are meaningful, relevant, and interesting.
- It provides thematic organization to items, such as through a story line or an episode.
- It offers tasks similar to real-world tasks.

Washback

Washback may have been called backwash, test impact, measurement-driven instruction, curriculum alignment, and test feedback (Brown & Hudson, 1998). Washback, according to Brown and Hudson, is "the effect of testing and assessment on the language teaching curriculum that is related to it" (p. 667). Likewise, washback is used to refer to the influence that a test has on teaching and learning (Hughes, 2003). Washback, therefore, can both be beneficial and detrimental or positive and negative. Positive washback takes place when the tests measure the same kinds of materials and skills stated in the objectives and taught in the courses (Brown &

Hudson, 1998). If a test encourages learning and teaching or if it provides opportunities for students and teachers to enhance the learning and teaching process, it is affecting language acquisition and instruction positively. Nonetheless, if the test causes too much anxiety for the students, teachers and parents, the waskback may be deemed as detrimental. Mismatches between the goals and objectives of the curriculum and tests can also be a source of negative washback. The example Brown and Hudson (ibid) gave is that "if a program sets a series of communicative performance objectives, but assesses the students at the end of the course with multiple-choice structure tests, a negative washback will probably begin to work against the students' being willing to cooperate in the curriculum and its objectives. Students soon spread the word about such mismatches, and they will generally insist on studying whatever is on the tests and will ignore any curriculum that is not directly related to it" (p. 668). Moreover, other examples of negative washback, as Mihai (2010) discussed, are when the teacher teaches to the test or when students cram for the test. In fact, cramming for a test or teaching to the test does not truly promote enduring learning, so the main purpose of education may largely be ignored.

Some Options in Language Assessment

The literature on testing is extensive, and it is beyond the scope of this paper to include all the types of language assessment. This section reviews some commonly used assessment methods seemingly useful to second language teachers in general. A definition of each type of assessment will be provided, and the strengths and weaknesses of each type of assessment will also be examined. Readers interested in specific test types for individual skills and areas including listening, speaking, reading, writing, grammar, and vocabulary can find them in Brown and Abeywickrama (2010). Coombe and Hubley (2003) noted that assessment in the field of English language teaching has come a long way from the days when it was simply equated with discrete-points, objective testing. They further added that although objective testing is still appropriate for certain purposes, assessment these days consists of a wide range of tools and techniques that range from testing an individual student's language ability to evaluating an entire program.

The method of assessment varies depending on the subjects and purposes of the assessment. Language testing is different from testing in other content areas as language teachers have more alternatives to make (Brown & Hudson, 1998). Brown and Hudson (ibid) identified three basic assessment types: (a) selected-response which includes true-false, matching, and multiple-choice assessments, (b) constructed response which includes fill-in, short answer, and performance assessments, and (c) personal-response which includes at least conference, portfolio, and self- and peer assessments.

A. Selected-Response

As Brown and Hudson (ibid) put it, selected-response assessments provide students with language material and ask them to select the correct answer among a limited set of choices. Because students do not usually produce any language in these assessments, they may work well for testing receptive skills such as reading and listening, as Brown and Hudson noted. Also, these authors commented that it can be relatively quick to administer these assessments and scoring them may be quite fast, easy, and objective. Nonetheless, Brown and Hudson noted that there are two main disadvantages in using these assessments.

- It is quite difficult to construct selected-response assessments.
- These assessments do not require students to produce language.

True-False

True-false is the type of assessment that requires students to choose either true or false to respond to the language sample given. The problem is that students have 50% chance of correct guessing, but if a large number of carefully constructed true-false items are employed, the overall score should overcome much of the guessing factor (Brown & Hudson, 1998). However, these authors also noted that if the language points the teacher wants to test lend themselves to two-way choices and enough number of items can be designed, true-false may be a good assessment method.

Matching

Matching requires students to match words, phrases, or sentences in one list to those in another. Whereas the advantages of matching is low guessing factor and the compact space needed, matching can only measure students' receptive knowledge of vocabulary (Brown & Hudson, 1998).

Multiple Choice

Multiple choice is the type of assessment that requires students to choose a correct answer among several options provided. Multiple-choice assessments have lower guessing factors than true-false, and they also are suitable for measuring a relatively wide variety of various kinds of precise learning points (Brown & Hudson, 1998). According to these authors, multiple-choice assessments are very efficient in testing reading, listening, grammar knowledge, and phoneme discrimination, as they can provide useful information about students' abilities and knowledge in such areas. Nevertheless, as Brown & Hudson pointed out, multiple choice assessments are often criticized because language use in real life is not multiple choice.

B. Constructed-Response Assessments

Whereas selected-response assessments are suitable for measuring receptive skills such as listening and reading and knowledge of vocabulary and grammar, constructed-response assessments are appropriate for productive skills such as speaking and writing. Constructed-response assessments require students to produce language through writing, speaking, or doing something else (Brown & Hudson, 1998). Moreover, these authors added that these assessments can be utilized to observe the interactions of receptive and productive skills such as the interaction of listening and speaking in an oral interview procedure or the interaction of reading and writing when students are required to read two academic articles and write an essay to compare and contrast them.

Performance

Performance assessments, as Brown and Hudson (1998) indicated, "require students to accomplish approximations of real-life, authentic tasks, usually using the productive skills of speaking or writing but also reading or writing or combining skills." Tasks used in these assessments may include essay writing, interview, problem-solving tasks, role playing, and pair and group discussions. Brown and Hudson (ibid, p. 662) pointed out three major requirements for performance assessments:

- Examinees are required to perform some sort of task.
- The tasks must be as authentic as possible.
- The performances are typically scored by qualified raters.

Brown and Hudson (ibid) also identified advantages and disadvantages of using performance assessments. In regard to advantages, performance assessments can elicit relatively authentic communication in testing situations. In terms of disadvantages, these assessments can be relatively difficult to construct and time-consuming to administer. Costs related to developing and administering performance assessments, rater training, rating sessions, score reporting are also a problem of considerable concern. Other problems with these assessments include (a) reliability such as inconsistencies among raters, subjectivity in scoring, and limited observations, (b) validity such as insufficient content coverage, lack of construct generalizability, the sensitivity of performance assessments to test method, task type, and scoring criteria, construct representation or problem of generalizing from a limited number of observations, and logistics issues such as collecting and storing audio or video files of performances, special equipment and security, and planning and administering scoring and construct-irrelevant variance (performance characteristics that have no relevance to students' real abilities). Finally, due to the limited number of prompts, students may be able to remember such prompts and pass them on to others, so they can prepare the responses to the prompts in advance, making it hard for the raters to determine the real level of language proficiency of the test-takers. In addition, because of the small number of prompts, teaching to the test is a possibility. Two specific performance assessment methods, interviews and essay tests, may deserve some attention.

Interviews

An interview in second language assessment is a method of assessing students' oral language proficiency by asking students to answers certain questions and the language students orally produce is used to determine their level of proficiency in oral communication. Interviews allow the interviewer or the classroom teacher to decide if the language produced is understandable, is used correctly in terms of vocabulary and grammar, and is an efficient vehicle for conveying the message the student wants to convey. However, as Brown and Abeywickrama (2010) indicated, the practicality of interviews is low as it is time-consuming. If time is not a constraint, classroom teachers may find interviews an authentic and relatively reliable and valid method to assess students' oral performance.

Essay Tests

For second language teachers, essay tests deserve significant attention, as they are frequently used in the classroom. An essay test can broadly be defined as a form of assessment in which students are required to respond to a question by composing a piece of writing such as an essay or a paragraph. In second language acquisition, essay tests may be regarded by many teachers one of the most reliable types of tests to evaluate student productive language use such as the use of vocabulary words and grammar structures to convey their ideas, opinions, or arguments. Moreover, students' ability to logically and clearly organize their writing can also be measured.

Reiner, Bothell, Sudweeks, and Wood (2002) observed that although essay tests are one of the most commonly employed methods of assessing student learning, many essay questions are poorly designed and ineffectively utilized. Below are some key considerations in writing essay questions (Hopkins, Stanley, & Hoptkins, 1990, pp.216-217):

1. Make definite provisions for preparing students for taking essay examinations.
2. Make sure that the questions are carefully focused.
3. Structure the content and length of questions.
4. Have a colleague review and critique the essay questions.
5. Avoid the use of optional questions, except when one is assessing writing ability where a choice of questions is desirable.
6. Restrict the use of the essay as an achievement test to those objectives for which it is best.

7. For general achievement testing, use several shorter questions rather than fewer longer questions.

Despite the common belief that essay tests are an excellent way to elicit learners' productive language and are a relatively reliable way to evaluate learners' ability to use written language, some limitations of essay tests may be of interest to classroom teachers, as they often have to evaluate learners' essay tests for classroom assessment purposes. Among the many problems with essay tests provided by Hopkins, Stanley, and Hopkins (1990), four serious limitations are worth mentioning: the halo effect, the item-to-item carryover effect, the test-to-test carryover effect, and the order effect.

First, the halo effect, the tendency to be influenced by other factors or characteristics when evaluating one specific characteristic of a person, may have an influence on the score given. For instance, when rating an essay written by a very hard-working, dedicated, and cooperative student, the teacher may subconsciously take all those positive characteristics of the student into consideration when giving a score to that essay. To eliminate this effect, rating essays anonymously is desirable and will guarantee more objective evaluation of students' essays.

Second, the item-to-item carryover effect refers to the situation when raters "acquire an impression of the student's knowledge on the initial item that "colors" their judgment of subsequent items" (Hopkins, Stanley, & Hopkins, 1990, p.201). To avoid this problem, teachers should be acutely aware that a response needs to be evaluated based on its own merits and should not be influenced by preceding questions on the test.

Third, the test-to-test carryover effect is the situation when the score of one paper is affected by the score of the preceding paper. Teachers may subconsciously compare the quality of the paper being graded with the one graded immediately before it. To achieve objective scoring, relying strictly on the rubric and comparing the essay being rated with the description of the rubric may ensure more objective and fair scoring.

Finally, the order effect refers to the situation when essays rated at the beginning of the scoring session receive higher scores than those at the end of the session. Hopkins, Stanley, and Hopkins (1990) suggested that raters may become weary and "in this physical and mental condition nothing looks quite as good as it otherwise might" (p. 202). This effect may be alleviated by taking frequent breaks after every one or two hours of scoring.

C. Personal-Response Assessments

Brown and Hudson (1998) indicated that personal-response assessments require students to produce language to communicate what they want to communicate. These assessments, as Brown and Hudson noted, are beneficial as they "provide personal or individualized assessment, can be directly related to and integrated in the curriculum, and can assess learning processes in an ongoing manner throughout the term of instruction" (p. 663). Nonetheless, the disadvantage of these assessments, as mentioned by Brown and Hudson, is that they are quite difficult to design, organize. and score objectively.

Conferences

Conference assessments occur when students are required to visit teachers' offices to discuss a particular piece of work or learning process, or both (Brown & Hudson, ibid). The benefits of these assessments are also mentioned by Brown and Hudson as follows:

- Teachers can use conference assessments to foster student reflection on their own learning processes.
- Teachers can use conference assessments to help students develop better self-images.

- Teachers can use conference assessments to elicit language performances on specific tasks, skills or language points.
- Teachers can use conference assessments to inform, observe, mold, and collect information about students.

The main drawbacks of these assessments, as provided by Brown and Hudson, are that they are quite subjective, difficult, and time-consuming to grade and they are usually not scored or rated.

Portfolios

Portfolio assessment is an ongoing process in which the student and teacher choose samples of student work to include in a collection, the purpose of which is showing the student's progress (Hancock, 1994). Hancock further indicated items that can go into a portfolio: samples of student creative work, tests, quizzes, homework, projects and assignments, audiotapes of oral work, student diary entries, log of work on a particular assignment, self-assessments, and comments from peers and teachers. Brown and Hudson (1998) discussed the advantages and disadvantages of portfolios at length, but the main points of their discussion may be summarized as follows.

As regards their advantages, portfolios can strengthen student learning, enhance the teacher's role, and improve testing processes. Five problems with using portfolio assessments are issues of design decisions (e.g., grading criteria, components of the portfolios…), logistical issues (e.g., time and resources needed for portfolio assessments), interpretation issues, reliability issues, and validity issues.

Self- and Peer Assessments

In self-assessments, as described by Brown and Hudson (1998), students have to rate their own language through performance self-assessments (students reading a situation and deciding how well they would respond in it), comprehension self-assessments (students reading a situation and deciding how well they would comprehend it), and observation self-assessments (students listening to audio or video recordings of their own language performance and deciding how well they think they have performed). Peer assessments, as the name suggests, involve students assess the language produced by their peers. Brown and Abeywickrama (2010) classified self- and peer- assessments into five categories: (a) direct assessments of a specific performance, (b) indirect assessment of general competence, (c) metacognitive assessment, (d) socioaffective assesement, and (e) student-generated tests (see Brown & Abeywickrama (2010) for more detailed explanations and examples of each of the assessment category). Brown and Hudson identified four important benefits of self-assessments.

- They can be developed to be administered relatively quickly.
- They always involve students in the assessment process.
- By involving students in the assessment process, they make students better understand what it means to learn a language autonomously.
- Both student involvement and their greater autonomy can greatly increase students' motivation to learn the target language.

Citing from Blanche (1988) and Yamashita (1996), Brown and Hudson (ibid) indicated that some disadvantages of self-assessments are as follows.

- The accuracy of self-assessments varies according to the linguistic skills and materials in the evaluations.
- Proficient language students have a tendency to underestimate their abilities.
- Scores from self-assessments may be affected by factors such as past academic records, career aspirations, peer-group or parental expectations, and lack of training in self study.

Concluding Remarks

The review of common assessment options discussed above has shown that the knowledge and skills needed for designing practical, authentic, reliable, and valid tests are likely to be real challenges for most classroom teachers who are seldom fully trained to construct quality tests. Teachers, nevertheless, usually have at their disposal a wide range of choices depending on the contexts where they work and the evaluation culture of the language program. If teachers are not charged with the responsibility of constructing tests for their own classes, they may be provided with tests to use for classroom evaluation. However, in cases where classroom teachers are required to produce their own tests to evaluate their students' learning outcomes and progress, it is critical that teachers are well informed of available tests to adopt or adapt, as test construction is usually an onerous task. One particular useful publication that is intended to help teachers create and analyze language tests is Carr (2011). Alternatively, instead of creating a test, classroom teachers may find it less challenging to use ready-made tests or test-generator CD-ROMs that accompany the textbooks they use for their class. The quality of tests accompanying textbooks may vary, but teachers can modify or adapt the tests to fit the assessment standards at their institutions. Regrettably, not all textbooks are accompanied by tests, which puts more responsibility on the teachers using them. Summative assessments such as final exams or tests may not always be the best way to evaluate students learning, as they inevitably put students through a great deal of stress and anxiety. One possible alternative to final tests is evaluating students during the course of their study through formative assessment, which may be a more learner-friendly method to assess student learning. A combination of both formative and summative assessments may also be a balanced approach to evaluating student progress, especially for second language learners, as no test can possibly measure all areas of skills and knowledge that learners have mastered. A balanced approach to evaluating second language students can ensure that the results of the assessments are more reliable, and students who do not perform well under pressure and stress during exams may find it a fairer method of second language assessment.

References

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

Bachman, L., & Palmer, A. S. (1996). *Language testing in practice*. New York, NY: Oxford University Press.

Blanch, P. (1988). *Self-assessment of foreign language skills: Implication for teachers and researchers*. RELC Journal, 19, 75-76.

Brown, H. D., & Abeywickrama, P. (2010) *Language assessment: principles and practices*. White Plains, NY: Pearson.

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly*, 32 (4), 653-675.

Carr, N. T. (2011). *Designing and analyzing language tests*. New York, NY: Oxford University Press.

Coombe, C. A., & Hubley, N. J. (2003). *Assessment Practices*. Alexandria, VA: TESOL.

Coombe, C., Folse, K., & Hubley, N. (2007). *A practical guide to assessing English language learners*. Ann Arbor: MI: The University of Michigan Press.

Hancock, C. R. (1994). *Alternative Assessment and second language study: What and why?* Center for Applied Linguistics. Accessed on 07/19/2012. URL: http://www.cal.org/resources/digest/hancoc01.html

Hopkins, K. D., Stanley, J. C., & Hopkins, B. R. (1990). *Educational and psychological measurement and evaluation*. Needham Heights, MA: Allyn & Bacon.

Hughes, A. (2003). *Testing for language teachers*. Cambridge, MA: Cambridge University Press.

Mihai, F. M. (2010). *Assessing English language learners in the content areas: A research-into-practice guide for educators*. Ann Arbor, MA: University of Michigan Press.

Passerini, K., & Granger, M. J. (2000). A developmental model for distance learning using the Internet. *Computer & Education*, 34, 1-15.

Richards, J. C. (2001). *Curriculum development in language teaching*. New York, NY: Cambridge University Press.

Reiner, C. M., Bothell, T. W., Sudweeks, & Wood, B. (2002). *Preparing effective essay questions: A self-directed workbook for educators*. Accessed on 08/06/2012. URL: http://testing.byu.edu/info/handbooks/WritingEffectiveEssayQuestions.pdf

Stoynoff, S., & Chapelle, C. A. (2005). *ESOL Tests and Testing*. Alexandria, VA: TESOL.

Yamashita, S. O. (1996). *Six measures of JSL pragmatics*. Honolulu, HI: University of Hawai'i Press.