A Review of the Literature on ESL Literacy Assessment

British Columbia Lower Mainland ESL Assessment Consortium

Dennis Murphy Odo

May 13, 2010

Table of Contents

Executive Summary

Scope and Purpose

This literature review was conducted to provide an overview of the major themes and findings in the research and professional literature for second language literacy assessment. The report should be useful to K-12 ESL teachers or school board assessment professionals who are seeking an overview of the professional and research literature on formal and informal approaches to English as a second language literacy assessment.

Methods

This review was conducted by consulting the research and professional literature on second language literacy assessment methods. Research studies were included if they were published in peer-reviewed journals. Books were included if they were published by reputable publishers within the area of second language education.

Historical Development of L2 Assessment

The first section of the report outlines the historical development of second language assessment methods. Some of the main findings are:

- During the pre-scientific era of language testing, teachers usually developed their own idiosyncratic and subjective assessments for their classes.
- Psychometric tests are more reliable and objective than their predecessors. Test questions are discrete and closed so that only one answer could be correct, thus, minimizing scorer subjectivity and giving the tests the appearance of being fair.
- Integrative tests include role-plays, interviews, problem-solving tests and so forth which encouraged test takers to integrate several skills at the same time to assess communicative competence.
- Performance-based assessments require actual performances of relevant tasks are required of candidates instead of abstract demonstrated through pencil and paper tests.

Current Approaches to Assessment

This section explains the features of norm referenced and criterion referenced assessments and discusses their relative strengths and weaknesses.

Norm-referenced

- Most standardized tests are norm referenced. They are used primarily to determine test takers' aptitude or general language proficiency to place them into a level within a language program.
- In essence, a norm referenced test assesses an individual test taker by comparing her score to the performance of others who have taken the same test.
- There has been some criticism of norm-referenced tests largely based on the way the results are being used for high-stakes assessments.

Criterion-referenced

- These types of tests are usually developed by the classroom teacher to measure student learning of the instructional objectives in the classroom.
- The score is interpreted based on criteria that were established beforehand instead of in relation to other test takers as with a norm-referenced test.
- It may be a challenge to decide a suitable minimum level of performance to pass a test. That is, there is no test percentage score that automatically tells the teacher that a learner is ready for the next level.

Alternative Assessments

Three of the most popular forms of alternative assessment are discussed. These are self/peer assessment, portfolio assessment and conferences. The relative strengths and weaknesses of each are explored. The research literature on each of their effectiveness for ESL students is also reviewed. Experts suggest alternative assessments as an solution to the danger of relying on a single traditional test score as the basis for passing a course because of the inherent error that can exist in tests.

Self/peer Assessment

- The main benefits of self and peer assessment are that it helps to develop students' meta-cognitive awareness of their own learning and involves them more in the assessment process.
- Research on peer assessments indicates that they are valid and reliable classroom assessment tools.
- Findings on self assessment are mixed. Some say that self assessments are generally accurate but others say they do not correspond to other forms of evaluation.

Portfolio Assessment

- Portfolio assessment attempts to provide a more authentic and all-encompassing measure of performance than traditional assessments.
- One significant advantage for portfolios is that they allow learners to use their teachers and peers as resources to help them improve their work.
- Studies have linked portfolio assessment to gains in reading comprehension and they have also been found to better predict students' future success than other standardized assessments.

Conferences

- During conference assessment, the student meets with the teacher to discuss his or her work or how his or her learning is proceeding.
- The unique aim of conferencing is that it focuses explicitly on learning processes and strategies for more efficient literacy learning.
- Empirical findings for the benefits of writing conferences for ESL learners have been mixed. Some have found that the personality of the learner had a significant bearing on the direction that writing conferences took. Others found that conferencing alone is not enough to encourage students to use a variety of discourse strategies such as sharing their opinions and asking for clarification when necessary.

Testing Techniques for Second Language Assessments

Popular second language literacy assessment techniques are reviewed next. This includes closed techniques such as multiple choice and cloze and its adaptations. Open ended testing techniques such as short response and essay questions are also discussed as is the use of rubrics for ESL students.

Research Findings for Multiple Choice Tests:

- The validity of multiple-choice tests is based on an unproven and possibly un-provable assumption that adding up the test takers' scores on a series of discrete items will provide a representative global measure of the test taker's language ability. Critics argue that language use is complex and contextualized so testing of language must be complex and contextualized.

- Developing valid and reliable multiple-choice test items is a time-consuming process that classroom teachers may be unable to do well.
- Students can exploit the nature of the multiple-choice test design to artificially inflate their test scores.

Research Findings for Cloze/ maze Tests:

- Several scholars have praised the cloze tests as a relatively easily developed and administered assessment for ESL learners.
- Some claim that the cloze best represents word or sentence level writing ability and that it does not really measure connected discourse. Therefore, it may not actually measure text comprehension but rather syntactic knowledge.
- Empirical results on the use of cloze tests to judge the reading comprehension ability of EAL students are divided.
- The maze procedure was found to be valid and reliable measure of content area reading in at least one study and learners were found to score better on maze test than on traditional cloze.
- A criticism of maze tests is that they are not as authentic because readers have to stop reading and try putting the choices in the blank to see which one is the most suitable word.

Informal Literacy Assessments for Different Grade Levels

Primary

The concept about print test was found to be effective with primary school learners. Research findings for concepts of print reveal:

- In one study, bilingual children were found to grasp the concepts of print identified better than monolingual children.
- Another study revealed that the structure of the first language is strongly related to a bilingual child's development of concepts of print.

Elementary

Informal Reading inventories (IRIs) were reported to be effective as elementary school assessments

- Much of the research into the validity and reliability of IRIs was conducted in the 1970s and 1980s. Therefore, the question remains whether the concerns raised at that time are still relevant today.
- IRIs are suitable for low stakes classroom decisions and materials selection but not for decisions with more serious consequences such as the detection of reading disabilities.
- IRIs are not reliable for measuring word recognition and comprehension skills in ESL students.
- L1 language transfer is an important element for the diagnosis of ESL students' reading ability and IRIs do not take into consideration potential L1 influences.

Secondary

Secondary school assessments that have proven to be useful are summaries, information transfer tasks (e.g., Venn diagrams), and various forms of vocabulary assessment.

- The use of summaries for second language assessment is under-researched. However, the studies that have been conducted affirm that teaching readers how to summarize aids their ability to effectively summarize.
- Research on the helpfulness of information transfer tasks reports that they do have potential for enhancing content area reading comprehension instruction. However, there does not appear to be a large research base underlying their efficacy as assessments.
- Language assessments that determine whether or not language learners can use vocabulary in meaningful context such as yes/no format, word associates format, and vocabulary knowledge scale tests are as valid and reliable as traditional vocabulary measures.

<u>Computer Based Testing</u>

Computer based assessment is becoming increasingly prevalent in the second language classroom. One type of computer based assessment frequently discussed in the literature is adaptive tests. Adaptive tests (CAT) promise to be more specified to test takers' level. However, in practice, theoretical and resource constraints have called into question the viability of the CAT in a wide number of contexts. Benefits of web based tests (WBT) such as user familiarity with the interface were also discussed. Research has shown web-based assessment to have comparable scores to paper-based assessment.

Strengths of CBT

- May allow for positive watch back because learners preparing for these sorts of second language assessments could develop their computer literacy in the process.
- CAT may offer improved test security because most test takers would be unable to memorize all of the questions for the test beforehand due to the large number of questions involved.
- CAT test takers are appropriately challenged because test questions are pitched to their ability level.
- WBT allows test takers to take assessments from the comfort of their own home anywhere in the world with a computer and access to the Internet.
- WBT test takers who have experience with using the Internet may actually find the web-based interface to be more recognizable than a traditional computer-based assessment.
- WBT assessment may be more inexpensive than traditional computer-based assessment because the tests could be used on pre-existing browsers instead of needing to develop software from scratch.

Potential Challenges with CBT

- There are issues with validity in that it is still unknown whether computer assessments measure the same abilities that traditional assessments measure.
- Concern that test takers from different nationalities do not have the same access to a computer and this may affect their test performance.
- Questions over whether or not the results of computer assessments are comparable with those of traditional paper-based assessments.
- Paper-based tests should never totally disappear because there is always the chance that technology may fail.
- Second language assessment scholars increasingly need knowledge of computer assessment in addition to principles of second language assessment.

<u>Social Impact of Assessment</u>

Criticisms of Societal Impact of Second Language Assessments are also reflected upon. The main criticisms discussed in the literature reviewed for this report were:

- Second language assessments often have a lot of power over the educational trajectory of students' lives because they often determine eligibility to enter and exit programs.
- Tests can systematically discriminate against particular groups through test bias in methods and materials.

Conclusion

- There is a wide variety of assessment options available to fit the unique needs of EAL learners at all grade levels.
- Hopefully, it has also allowed for some reflection on the serious consequences that assessment has on learners' lives and the burden that brings.
- There is still a need for research in second language literacy assessment. In particular, there is a dearth of research into the efficacy of newer adaptations of the cloze method with second language learners or forms of informal classroom assessment such as using graphic organizers and communicative vocabulary assessment tasks.

Introduction

The number of speakers of English as an additional language In Canada continues to grow. In fact, according to Statistics Canada, by 2031 (CBC, 2010), one in four Canadians will have been born in another country. The British Columbia Ministry of Education reports that in 2008-2009, 22% of public school students in B.C. spoke a language other than English at home. This is a large influx of English language learners into our schools and it does not look like this trend will be changing in the near future. Since we have agreed to welcome these fellow human beings to Canada, we are responsible for helping them to succeed in our society (Oikonomidoy, 2007). However, the alarming disappearance rates of English-as-an-additional-language (EAL) learners from schools across Canada (Toronto, 53%) (Radwanksi, 1987), (Calgary, 73%) (Watt & Roessingh, 2001) and in the Lower Mainland (60%, Gunderson, 2007) tells us that more needs to be done.

Gunderson (2007) concluded that if we want to retain immigrant learners in school and prevent them from giving up and leaving, we must begin by teaching them English language and literacy skills that will help them to access the many texts that they are expected to learn from at school. A large part of ensuring that immigrant students are meeting the learning targets set for them is putting appropriate and informative assessments in place. These assessments must be valid and reliable because they often have serious consequences for learners' education and lives (Shohamy, 2000).

The British Columbia Lower Mainland ESL Assessment Consortium was formed in part to share information about successful local assessment practices to ensure that teachers of English as an Additional Language (EAL) learners would have access to appropriate assessments.

The purpose of this report is to survey the professional and research literature to learn about the literacy assessment practices that have been found to be effective for second language learners.

This review begins by tracing the evolution of approaches to second language assessment. Next, the features of norm-referenced and criterion-referenced tests are discussed in relation to their appropriate uses within a language program. From there, common open-ended, closed and alternative assessment tasks are evaluated in light of findings in the empirical research literature. Informal EAL literacy assessments used at the primary, elementary and secondary levels are reviewed. The section on computer adaptive testing explores some of the reasons why it has not become more widely used in second language assessment. The final section contains a brief discussion of the criticisms of the uses of modern tests.

Historical Development of Second Language Tests

Pre-scientific

During the pre-scientific era of language testing, the tests developed for language teaching were usually made by teachers for their own classes. They often consisted of asking students to translate passages to demonstrate their understanding. These tests were not necessarily valid or reliable because they were subjectively scored based on the teacher's own personal criteria. The inherent unfairness of these assessments eventually became apparent and steps were taken to rectify these disparities (Brown, 2005).

Psychometric

Baker (1989) argues that the first major advancement in the field of language testing came about shortly after the Second World War. This advancement eventually came to be known as psychometric testing. There were two main causes for its development. The first was the

refinement of large-scale psychological testing and more sophisticated statistical techniques for analyzing these types of tests. The second major cause of this shift in assessment was structural linguistics. Structural linguistics enabled language to be broken down into its smallest pieces and offered a description for how these pieces could be reassembled. This allowed testers to be able to disassemble language in order to teach and test it more systematically. This form of assessment provided a significant advantage over pre-scientific tests. In many respects, this explains why psychometric assessment remains popular. They have features that make them more reliable and objective than their predecessors. Test questions are discrete and closed so that only one answer could be correct, thus, minimizing scorer subjectivity and giving the tests the appearance of being fair.

Communicative/ Integrative

In due course, shortcomings in psychometric tests became apparent. Although they seemed reliable and objective, they were limited by a focus on discrete pieces of language and not the language as a whole. However, in the 1970s and 1980s scholars were beginning to realize that language was more than a collection of discrete pieces. That is, language was more than the sum of its parts. Thus, they argued that communicative competence needed to be accounted for in language tests so new forms of assessment were developed. These newer assessments included role-plays, interviews, problem-solving tests and so forth which encouraged test takers to integrate several skills at the same time (Brown, 2005).

Performance/ Task based

Another more recent development in second language assessment is performance or task-based assessment. According to McNamara (1996), task-based assessment began in the 1960s because of the practical needs of English for specific purposes (ESP) courses. The need for

learners taking ESP courses to demonstrate their competence in a specialized area of language use led to these more practically-oriented forms of assessment.

McNamara notes that the main characteristic of performance assessments is that test takers actually have to complete similar tasks that they would be expected to do in the target context instead of showing their knowledge through pencil and paper tests (Slater, 1980, cited in McNamara, 1996). compared traditional and performance based assessments designed for emergency medical technicians. The traditional assessment might have the test taker complete a multiple choice test of knowledge of procedures for dealing with an accidental poisoning. In contrast, the performance assessment would incorporate a simulation where the test taker had to interact with a parent whose child had accidentally ingested poison. In this way, the assessment demonstrates that the test taker is able to use the language in more authentic situations. Task authenticity is the most essential aspect of performance task design (Bailey, 1998).

Current Approaches to Second Language Assessment

Norm-referenced

Two main types of language tests are currently available to teachers and each of these types is used for different purposes. One type is known as norm referenced. Most standardized tests are norm referenced. Norm referenced tests are used primarily to determine test takers' aptitude or general language proficiency to place them into a level within a language program. Since these tests are a measure of general language ability, test takers do not normally know what the specific content of the test will be before they take it (Brown, 2005). These tests are often used with large numbers of test takers to screen out people from selective programs. For

example, the TOEFL is often used to screen foreign students entering American universities (Bailey, 1998).

In essence, a norm referenced test assesses an individual test taker by comparing her score to the performance of others who have taken the same test. In order for the statistical assumptions that the test is based on to be satisfied, the norming group that the test taker is being compared to must be similar to the test taker on characteristics measured by the test. These tests can be normed by comparing the test taker with the group that she is taking the test with or another different group of test takers (Bachman, 1990). Scores for norm referenced tests are given as percentiles. These differ from percentages because they tell the test taker her place in a particular distribution. For instance, if she is in the 96th percentile then she performed better than 96 percent of the people who took the test (Bailey, 1998). Since the aim of the test is to make distinctions between test takers, the questions are designed to spread test takers out along a continuum of scores (Brown, 2005).

There has been some criticism of norm-referenced tests largely based on the way the results are being used. One common misuse is their frequent use for high-stakes assessments. That is, they are often used as gatekeepers. Neill and Medina (1989) note that "from preschool to college, [standardized tests] have become the major criteria for a wide range of school decisions. Test scores limit programs that students can enter and dictate where students are placed." (p. 688) Learner performs will often dictate what future educational opportunities she has. A related criticism is that school personnel often use results to discriminate against children whose scores might reflect negatively on a school in district comparisons. Law and Eckes (1995) assert that

> Schools have been known to place many ESL students in Special Ed; keep them in federally funded compensatory programs such as Title VII and Chapter 1 long after they should have been exited into the mainstream; not test; or worst of all, encouraged him to drop out -- because of the effect the students will have on overall scores." (p.34)

A third shortcoming of these tests is that there is always a very real danger of teachers teaching to the test. That is, teachers teach the material and test taking strategies that they believe will help their students pass the test. Worthen (1993) observes that "educators were quick to realize that targeting their instruction at the specific knowledge and skills to be tested would yield a jump in test scores." (p. 446) There is always the risk that quality instruction will be sacrificed for practice with tricks and techniques that may help learners perform better on the test. This is a particular concern in an era of increased accountability.

Criterion-referenced

The second approach is known as criterion referenced. These types of tests are usually developed by the classroom teacher to measure student learning of the instructional objectives in the classroom (Brown, 2005). The main difference between criterion referenced tests and norm referenced is that scores on these tests are interpreted based on criteria that were established beforehand instead of in relationship to other test takers as with a norm-referenced test (Bailey, 1998). The distribution of scores for these tests cannot be expected to be normal because if all the learners know the material then they can all achieve a perfect score. This is not possible with a norm-referenced test (Brown, 2005). The best feature is that they allow evaluators to see how much knowledge test takers have instead of how much they have in comparison to each other (Bachman & Palmer, 1996). Additionally, unlike a norm-referenced test, the content of criterion-referenced tests is based on the curriculum so learners should have a good idea of what kinds of questions to expect.

Bailey (1998) claims that criteria-based assessments are more appropriate for classroom use than norm-based assessments. She points out that we use criteria-referenced tests to know if students have mastered a particular skill or an amount of course material. In addition, because these are classroom assessments, they are not used to screen students out. Instead, they are used for diagnosis for remediation and as measures of student achievement to inform teachers about who is ready to move to the next unit or level.

Critics claim that these measures also have drawbacks. Baker (1989), for instance, points out that it may be a challenge to decide a suitable minimum level of performance to pass a criterion-referenced test. That is, there is no test percentage score that automatically tells the teacher that a learner is ready for the next level. Of course, an answer to this critique is that the teacher should only be using these tests as one indicator among many to decide whether a learner is ready for the next level. Used this way, they should provide valuable information of learners' knowledge of curricular objectives. Brown (2005) raises another criticism that is normally reserved for norm referenced tests. That is the danger of teaching to the test. He points out that, unlike norm-referenced tests, teaching to the test is a good idea if the test reflects sound objectives that meet the learners' needs. In fact, teaching to this sort of test will ensure that teachers are following the objectives of the course.

A final point about these testing methods is that neither is inherently better than the other. They both serve valuable purposes to help teachers and administrators make decisions about students' language performance. The key is to know which test is best suited for a particular decision such as placement or diagnosis.

Alternative Assessments

Bailey (1998) points out that relying on a single traditional test score as the basis for passing a course can be dangerous for learners because of the inherent error that can exist in tests. She also notes that teachers are increasingly questioning the authenticity of traditional forms of testing as measures of learner capability. This has led them to explore alternate possibilities for language assessment. Drawing on the work of several other scholars, Brown and Hudson (1998) devised a list that includes twelve of the main features that the various forms of alternative assessment have in common. Some of the shared features are that they engage learners' higher-level, creative thinking, they require real-world tasks and provide feedback on learners' strengths and weaknesses. In the following section, three of the more popular forms of alternative assessment – self and peer assessment, portfolios and conferences – will be discussed.

Self and Peer Assessment

Brown (2005) defines self assessments as "any items wherein students are asked to rate their own knowledge, skills, or performances." (p. 58-59). He explains that their purpose is to give the teacher some insight into how the learners see their language ability developing. The goal of peer-assessment is to have students evaluate each other's work. Aebersold and Field (1997) note that the main benefits of self and peer assessment are that it helps to develop students' meta-cognitive awareness of their own learning and involves them more in the assessment process. Peer assessment also offers the advantage that learners may actually try harder during group work if they know that they are being evaluated by their peers. The disadvantage for both forms of assessment is that other stakeholders such as administrators or parents are often doubtful of their legitimacy.

Research on self assessments indicates that they are valid and reliable classroom assessment tools. After conducting a traditional literature review method, Blanche and Merino (1989) reported that there was an emerging pattern of overall agreement between self assessments and ratings based on other types external assessment criteria. They also noted that the accuracy of most students' self-estimates often depended on their own language ability and the design and content of the evaluations. Ross's (1998) meta-analysis of the literature related to self assessment of the four skills revealed that, for reading self assessment, learners' self-assessments strongly correlated with their actual reading performance. This supports Blanche and Merino's conclusion that self assessments are generally accurate.

There are some contrary findings such as Kwan and Leung (1996), and Orsmond et al. (1997) who observed that self and peer assessments do not correspond to other forms of evaluation. Kwan and Leung (1996) investigated the comparability of postsecondary students' peer assessment with that of their tutors. They concluded that the grades given to each of these groups were the same less than 50% of the time. Orsmond et al. (1997) studied first-year undergraduate biology students and found that there was a discrepancy in overall grades between self-evaluations and instructor evaluations. It should be noted, however, that both of these studies were with native speaking populations.

Portfolio Assessment

Cohen (1994) describes portfolio evaluation as "a system of assessment intended to help teachers get a more natural and prolonged assessment of students than through traditional means of assessment" (p. 336). Portfolios can be used for reading or writing. They are used for writing in order to give learners another option that may alleviate some of the pressure to perform found

in traditional writing tests. They are also used to collect in-progress samples of students' reading texts and their responses to those texts (Cohen, 1994). Portfolios usually contain information that can be used for formative and summative assessment. This could include tests, quizzes, reports, running records, teacher or peer feedback, journals and other creative student work (Afflerbach, 2007; Flippo, 2003). This work can be collected by the teacher, the student or both (Caldwell, 2002). Cohen (1994) notes several advantages of portfolios. One significant advantage is that they allow learners to use their teachers and peers as resources to help them improve their work. They can also help students to critically evaluate their work as they decide what they would like to keep as examples of their best work. One of the biggest problems with portfolios is that organizing and helping students decide what to keep in the portfolio can be time consuming. Another problem is that, if portfolios are judged by a team of teachers, a bad performance can reflect badly on the teacher whose student is being assessed and cause that teacher unnecessary stress (Cohen, 1994).

Research with sixth-grade ESL learners whose teacher used a portfolio assessment to evaluate them over the course of a school year revealed that they showed greater increases in their written fluency as measured by dialog journal word counts when compared with peers who did not use portfolios. There was also a significant gain in reading comprehension as measured by a cloze test (Newman et al., 1996). Another study of writing portfolios with EAL college freshmen concluded that portfolios were better at predicting students' future success than other standardized assessments (Song & August, 2002).

Conference Assessment

During conference assessment, the student meets with the teacher to discuss his or her work or how his or her learning is proceeding. This is a time when the teacher listens to the student's ideas and concerns about his or her literacy learning (Brown & Hudson, 1998). Flippo (2003) explains that the purpose of a conference is to give the teacher and the learner an opportunity to assess the learner's progress toward learning goals since the last conference. This is also the time when the teacher can provide encouragement and feedback to the learner. The aim of a conference is unique compared to other forms of assessment because it focuses explicitly on learning processes and strategies for more efficient literacy learning (Genesee & Upshur, 1996). Flippo (2003) insists that regular discussions with learners in the form of conferences are crucial to engage in ongoing monitoring of their progress toward literacy learning goals. Brown (2005) adds that conferences allow learners to draw on their teacher's knowledge and insights to identify the reading and writing strategies that they are able to use the most productively. The disadvantages of conferences are that "they are relatively time-consuming, difficult and subjective to grade, and typically not scored or rated at all." (Brown & Hudson, 1998, p. 662)

Empirical findings for the benefits of writing conferences for ESL learners have been mixed. One the one hand, Goldstein and Conrad (1990) found that the personality of the learner had a significant bearing on the direction that writing conferences took. Teachers often adjusted to the speaking style of the students which gave some students more opportunities to contribute to the dialog than others. This largely confirmed earlier results of Freedman and Sperling (1985). One contrary finding indicated that conferencing alone is not enough to encourage students to

use a variety of discourse strategies such as sharing their opinions and asking for clarification when necessary (Zamel, 1985).

Testing Techniques for Second Language Assessments

Two broad categories of questions that are used in second language assessment are closed questions and open-ended questions. Closed-questions provide test takers with options from which to choose and typically have only one correct answer. Answers to open-ended questions are constructed by the test taker and so there is much more variability possible.

Closed Techniques

A representative example of a closed question is a multiple choice question (Baker, 1989). These are most commonly used to assess listening and reading comprehension ability. Bailey (1998) notes that there are several reasons why teachers, schools, and assessment organizations prefer to use multiple choice tests: they are economical, they can be scored objectively (and thus appear to be fairer), and their appearance matches traditional notions of what test should look like. Baker (1989) explains that another attraction is that they consume relatively few human resources because the test scorers do not need any specialized knowledge besides the ability to follow scoring procedures. He adds that because test items are discrete and unrelated to each other, when an individual item is found to be unsatisfactory through piloting, it can be removed without affecting other items on the test (Baker, 1989).

Though multiple choice tests are appealing for economical and logistical reasons, they also have serious drawbacks. For instance, Brown (2005) indicates that the validity of multiple-choice tests is based on an unproven and possibly unprovable assumption. That is, adding up the test takers' score on a series of discrete items will provide a representative global measure of the

test taker's language ability. Critics disagree and maintain that discrete-item tests are too simplistic. They argue that language use is complex and contextualized so testing of language must be complex and contextualized.

A second point is that developing valid and reliable multiple-choice test items is a time-consuming process (Baker, 1989). Hughes (1989) claims that

> Good multiple-choice items are notoriously difficult to write. A great deal of time and effort has to go into their construction. Too many multiple-choice tests are written where such care and attention is not given (and indeed may not be possible). The result is a set of core items that cannot possibly provide accurate measurements." (p. 3)

Considering the amount of skill and effort required to write suitable items, one cannot help but question the results when busy, untrained teachers are asked to develop these sorts of questions.

A final insightful criticism of multiple-choice assessments comes from Alderson and his colleagues (1995). They point out that students can exploit the nature of the multiple-choice test design to artificially inflate their test scores. They are able to do this by using tricks that help them guess the correct answer by skilfully eliminating distracters. This results in an assessment of test-wiseness rather than language proficiency.

Cloze

A second type of popular ESL literacy assessment is the cloze test. Cloze passages are short texts that have one word deleted every *nth* number of words. They are used to measure test takers' reading skills interactively. Test takers are supposed to use textual cues in combination with their background knowledge to help them complete the cloze task (Cohen, 1994). According to Oller (1979), the cloze is based on the concept of pragmatic expectancy grammar.

This is a form of language competence that allows us to make predictions about language input which helps us to successfully process spoken or written text that is not completely clear. Bailey (1998) explains that two kinds of mental competence are drawn on to complete a cloze task. One is syntagmatic competence which tells the reader the part of speech of the word being read. The second is paradigmatic competence. This tells us the necessary semantic features the appropriate word must have to complete the blank. Some factors that affect the difficulty of a cloze passage are: the length of the text, amount of time allowed to complete the test, learners' familiarity with topic, genre, vocabulary and syntax, and the number of blanks (Bailey, 1998).

Several scholars have praised the cloze tests as a relatively easily developed and administered assessment for ESL learners. Cohen (1994) discusses its versatility noting that it has been used to measure readability, global reading skills, and grammar. Its supporters claim that it is also a suitable measure for listening comprehension and speaking. Hurley and Tinajero (2001) recommend the cloze because "first, it is flexible enough to allow teachers to assess large groups of students at the same time or assess students individually. Second, information gathered in the assessment procedure can be used to select reading material for students that is challenging but not frustrating" (p. 19). Similarly, Law and Eckes (1995) contend that "close tests provide a window into the strategies the student is using to gain meaning, as well as insight into how sophisticated his or her skill level in English is." (p. 68-69)

Not all of the commentary on cloze tests has been favourable. Markham (1985), for instance, claims that the cloze best represents word or sentence level writing ability and that it does not really measure connected discourse. Another fault pointed out by Cohen (1994) is that it

might be possible to successfully complete a cloze task without understanding the meaning of the text. That is, it may not actually measure text comprehension but rather syntactic knowledge.

A third criticism from Alderson et al. (1995) was that the choice of the first word in the close passage can dramatically influence the validity and reliability of the test. Consequently, the resulting tests can be very different depending on which word is deleted at the beginning because the subsequent deletions can be mostly function words which are relatively easy to restore or content words which are significantly more difficult to retrieve.

Findings of empirical research on the use of cloze tests to judge the reading comprehension ability of ESL students are divided. Some studies support the validity of cloze tests as a reading comprehension test that provides clear evidence of readers' ability to process text at the intersentential level. That is, to read text as coherent connected discourse (Chavez-Oller *et al*., 1994; McKenna & Layton, 1990; Oller & Jonz, 1994). Others have found that cloze tests measure only the ability to use local syntactic constraints and not contextual information. Thus, readers were found only to process structures at the sentence level (Shanahan *et al*., 1982; Markman, 1985; Abraham & Chapelle, 1992).

Cloze Adaptations

In response to the criticisms discussed above, other variations on cloze tasks have been developed or adapted for EAL assessment. One such adaptation is the rational cloze or gap-fill. This is like a traditional cloze except that instead of words being deleted at regular intervals they are deleted based on the test creators judgement. Alderson et al. (1995) recommend using this exercise as opposed to the cloze test because the test creator has more control over which words to blank out. This provides the test developer more flexibility to choose words that will better

measure the examinee's reading comprehension ability. Empirical research on rational cloze has shown it to be an effective form of literacy assessment (Bachman 1982; Yamashita, 2003).

In order to test whether or not the word choice for deletions in cloze affected the validity of the test, Bachman (1982) devised a rational cloze passage that included deletions that were designed to assess readers knowledge of syntactic and cohesion features of the text. His findings were that rational deletions in a text effectively measured test takers' knowledge of syntactic and discourse relationships in the text or text comprehension beyond the sentence level. These findings corresponded with the results of a more recent study by Yamashita (2003) who compared the standard cloze technique with rational cloze. He concluded that the rational cloze better assessed text-level processing and was better able to accurately differentiate between skilled and less skilled readers.

Another variation on the cloze is the C-test. This is a type of cloze test that has individual letters blanked out of the second half of every second word instead of entire words (e.g., elep _ _ _ _). The benefit of using the C-test is that it forces discourse-level processing because learners have more information about the words used. The main drawback of C-tests is that – like most forms of cloze – students do not normally like them. This is particularly the case with c-tests because learners have to fill in many more blanks (Bailey, 1998).

Research into the validity and reliability of c-tests has shown that they are reasonably valid and reliable tests (Dornyei & Katona, 1992). It was also shown to be a valid measure of both micro (i.e., utterance) and macro (i.e., discourse) aspects of language knowledge (Babaii & Ansary, 2001). Nevertheless, this finding is not without some controversy because other researchers contend that "while c-testing may be a legitimate device for L1 testing, it lacks a

theoretical basis for application with FL learners" (McBeath, 1989, p. 36) and that they are not good measures of general proficiency for lower level learners (Carrol, 1986).

A third variation is the maze procedure. The maze procedure is a type of cloze test except that instead of having words blanked at regular intervals learners are given several choices to complete the blank. The advantage of the maze procedure is that the multiple choice format limits the possibility of unexpected or ambiguous answers. Research has shown that the optimal choice of distracter depends on the level of the test taker. The best types of distracters for beginners are semantically or grammatically incorrect. For intermediate learners, semantically incorrect distracters are best. Distracters with problems in style or register were found to be the most suitable for advanced learners.

The maze procedure was found to be valid and reliable measure of content area reading in at least one study (Espin & Foegen, 1996). As well, learners were found to score better on maze test than on traditional cloze (Abraham & Chapelle, 1992). Nevertheless, the maze procedure has some critics as well. Steinman (2002) contends that while maze tests are easy to score for teachers, they are not as authentic because readers have to stop reading and try putting the choices in the blank to see which one is the most suitable word. Of course, this concern about authenticity could just as easily be expressed about any cloze test and most other forms of reading assessment. One final criticism was raised by Propst and Baldauf (1979) who noted that "multiple-choice cloze test construction is not an easy, mechanical task which all teachers can do, but requires sophisticated knowledge about option selection" and for this reason, they suggest "matching cloze." (p. 685). They suggest teachers develop "matching cloze" tasks that allow students to select the correct word from an alphabetized list.

Open-ended Test Items

Short Response

Several types of open-ended test items are commonly used for assessing second language performance. The short response test item is one that test takers can answer using a range of responses from a few phrases to a few sentences (Brown, 2005). Two main guidelines for these items are that they should be a clear and direct question and they show require only one short answer. The strength of this type of item is that they can provide a short targeted answer that lets test takers relate their understanding of a concept or process in their own words. The main problem with this sort of item is the same one that plagues other open-ended test items – subjectivity. The teacher will have to make subjective judgements about the quality of a test taker's answers relative to her peers (Brown, 2005).

Essay Response

Another test item that requires a more extended answer is the essay response. These types of responses are most commonly used to assess writing ability. Test takers are typically required to respond to some sort of prompt. Traditionally, this type of test question was used to assess learners' writing ability with an emphasis on grammatical form though at present most test developers acknowledge the need to focus on both form and content (Kern, 2000). A positive feature is that it gives learners the opportunity to use many of their linguistic resources creatively to express themselves on a particular topic. A serious shortcoming is that while the pedagogy for ESL writing instruction has evolved this assessment form has not. The pedagogy has moved from focussing on a final product to paying more attention to the writing process (i.e., initial idea generation, drafting and revision) (Camp, 1993). The assessment of writing has not paralleled these changes in pedagogy. The typical essay on the majority of high-stakes measures still asks

the test taker to produce a polished product in 50 minutes with minimal time to reflect, consult with peers or revise their work (Lee, 2006). This approach is still clearly product-focussed.

Rubrics

A common tool used to evaluate the essay response is the rubric. In theory, any form of creative open-ended assessment can be evaluated using a rubric but it seems that they are most commonly used with essay tests (Bailey, 1998). Bailey (1998) explains that the scoring criteria used in a rubric can either "identify skills that are demonstrated (or not) [or] they can also identify the extent to which skill is demonstrated" (p. 187). There are two approaches to scoring rubrics. When "teachers use a single general scale to give a single global rating for each student's language production" (Brown, 2005, p. 54) they are using a holistic approach. A useful feature is that there tends to be higher agreement among raters. A shortcoming is that they do not allow for detailed feedback to test takers. The analytic approach has “teachers rate various aspects of each student's language production separately” (Brown, 2005, p. 54). Its advantage is that it allows for more detailed feedback to be given to test takers but it also tends to have less inter-rater agreement.

Reliability research is mixed. Weigle (2002) found that training helps raters to come to a temporary agreement about the standards to be applied in a given scoring session, but raters are never in total agreement. If the same raters re-marked the same essays at different times, many essays received somewhat different scores (Weir, 2005). This result calls the intra-rater reliability of these measures into question. These conclusions were supported by other research that found scoring to be a major contributor to measurement error in holistic assessments (Cherry & Meyer, 1993).

Other research supports the reliability of rubrics. A contrasting finding of one case study was that both analytic and holistic rubrics can be used with a high degree of reliability. This high degree of reliability was found when holistic rubrics were compared with each other or when they were compared with analytic rubrics. Another interesting result from this study was the discovery of raters' preference to use analytical rubrics (East & Young, 2007).

Classroom Literacy Assessments for Different Grade Levels

In the following section, various assessment techniques used with EAL students from K through 12 are discussed.

Primary School Assessments

Alderson (2000) notes that the main skills that need to be developed in emergent literacy assessment for English speaking children are "sound-symbol correspondences … knowledge of the alphabet and the conventions of print, and their word recognition ability increases in terms of number of words recognized, and speed and accuracy of recognition" (p. 275). These skills can be developed by using early literacy assessment to ensure that these essential factors are in place when the child enters school (Caldwell, 2002).

Pre-awareness

According to Law and Eckles (1995), the second language reader progresses through a number of stages that are comparable to those of the first language reader. The first of these is pre-awareness. This stage is characterized by a complete lack of knowledge of what texts are or what they are used for. At this juncture, assessment takes the form of evaluating oral language strengths and needs. This can be achieved through using formal assessments that are administered by appropriately trained personnel (Caldwell, 2002). An example of such an

assessment is the Pre-IPT Oral English Language Proficiency Test. Teachers should also assess students engaging them in dialog and listening to them interact with their peers (Caldwell, 2002).

Emerging Awareness

Emerging awareness is the next stage. At this point, learners begin to understand that print represents a message. A suitable assessment during this stage may be the Concepts about Print (CAP) test. This test evaluates whether the child can recognize the front of a book, the beginning of the text, that the print carries a message and so forth (Caldwell, 2002). An alternative more informal assessment is to have the teacher simply share a book with the child. The teacher will be able to learn a lot from the child by asking her questions and watching her behaviour as she reads.

After reviewing the literature related to English language learners who are also struggling readers, Klingner, Artiles, & Barletta, (2006) concluded that print awareness is related to second language ability for learners of English as a second language. Bialystok et al. (2000) found that, in general, bilingual children grasped the concepts of print in the study better than monolingual children. They also concluded that the structure of the first language is strongly related to a bilingual child's development of concepts of print. That is, Chinese-English bilinguals had more difficulty with English concepts than Hebrew-English bilinguals because Chinese is a non-alphabetic script and Hebrew is alphabetic. The wide discrepancy between their two languages confuses Chinese-English bilinguals and thus made it more difficult for them to grasp the concepts of print. Kalia (2007) and Kalia and Reese (2009) concluded that exposure to book reading in English with parents at home is closely associated with bilingual children’s oral language, narrative and literacy development in their second language. Measures of story book

knowledge were found to predict bilingual children's oral English language skills and knowledge of concepts about print.

Two more concepts that learners acquire during their emerging awareness are alphabetic knowledge and phonological and phonemic awareness. One relatively straightforward way of evaluating alphabetic knowledge suggested by Caldwell (2002) is to type the upper and lower case letters in random order on a piece of paper and ask the learner to identify the letter and say what sound it makes. Phonological awareness is the knowledge that language is made up of words which are made up of syllables and morphemes which are comprised of individual phonemes. This understanding that words are made up of individual sounds is phonemic awareness (Flippo, 2003). Suggested examples of tasks where children could demonstrate both phonemic and phonological awareness are identifying rhyming words or verbally blending or segmenting the individual sounds in words (Flippo, 2003).

Elementary School Assessments

Emerging Reading

Law and Eckles (1995) state that the next stage in developing literacy – emerging reading – tends to progress in elementary school. This is when learners begin to focus on the individual words of the text as they read and are able to identify some individual words in the text. One method for assessing this feature is through the use of miscue analysis and running records. The task of miscue analysis is to scrutinize learners' adding, changing, replacing or omitting words as they read aloud with the goal of understanding what strategies readers use for word recognition (Flippo, 2003). These miscues are then analyzed for their "graphemic, phonemic, morphological, syntactic, and semantic similarity with expected responses" (Alderson, 2000, p. 341). Running

records are similar to miscue analysis in the way that readers' errors are recorded and analyzed as they read aloud. The main difference is that running records rely on checkmarks for words that are correctly read and notations for errors while miscue analyses are usually recorded on a copy of the text and a miscue analysis coding form (Flippo, 2003).

Gaining Fluency

As they progress, learners begin to gain fluency in reading. This phase is distinguished by learners' more independent and silent reading (Law & Eckles, 1995). Caldwell (2002) recommends several methods for assessing the components of fluency. The first component is reading speed or rate which he suggests can be measured in words per minute. He points out that reading rates will certainly vary with texts and reading purposes. He contends that reading rate is probably best for identifying students who are extremely slow readers at their independent and instructional levels so that they can be given supports. A second method for assessing fluency is to given a student a list of sight words and note the number of words it takes her longer than one second to read. If it takes her longer than a second to read more than half of the words she may need remediation for sight vocabulary. Lastly, a checklist can be developed to assess students' verbal expression during oral reading.

According to Afflerbach (2007), informal reading inventories (IRI) share some features with running records. They both attempt to identify how readers' decode and create meaning using texts. They also both try to use information gained from analysis of readers' miscues to identify their strengths and weaknesses to improve future reading instruction. Gunderson (2009) explains that IRI's consist of graded word lists and graded comprehension passages. Teachers are supposed to compare learners' performance on reading the word lists and reading passages with

their performance on the comprehension questions to get a rough indication of their reading level. The reading levels used for IRIs are independent, instructional and frustration. Independent level texts can be read autonomously, instructional texts require some assistance and frustration level are too difficult, even with support. Gunderson (2009) also points out that IRI's are best made up of materials that the learners will actually be expected to read.

McKenna (1983) reviewed the empirical research for IRIs and identified several problems. Limitations with the passages included difficulty selecting passages that represented the book's difficulty level. A second problem was the powerful effect of familiarity with passage content on student performance. Two reading related problems were the accuracy of judging miscues and determining whether silent or oral reading scores are better representations of comprehension. A shortcoming of the comprehension questions is that learners are often able to answer them without actually having to understand the text. There was also the questionable validity of using decontextualized word lists as indicators of word recognition ability. Lastly, he mentioned the problem of using scoring criteria that do not have a research basis for their validity.

In a more recent literature review, Specter (2005) points out that much of the research into the validity and reliability of IRI's was conducted in the 1970s and 1980s. The question remains whether the concerns raised at that time are still relevant today. Specter's investigation of commercially available IRIs revealed that most publishers do not report any information on their reliability. Based on her analysis, she concluded that IRIs are suitable for low stakes classroom decisions and materials selection but not for decisions with more serious consequences such as the detection of reading disabilities.

Another dissertation study that probed the use of IRIs with ESL students found that they could in fact be useful. However, modifications to administration and scoring procedures had to be made to adjust for language transfer and second language development. Centurion (1985) observed two noteworthy problems with the use of IRI's for EAL learners. First, IRIs are not reliable for measuring word recognition and comprehension skills in EAL students. Second, L1 language transfer is an important element for the diagnosis of EAL students' reading ability and IRIs do not take into consideration potential L1 influences.

Increasing Fluency

The next phase is the fluency stage (Law & Eckles, 1995). This is when the learner can begin to read unfamiliar texts with some confidence and draw inferences from texts. In addition to the other assessments discussed above, another potentially useful tool might be interest assessments. They are simply a list of questions in an informal survey that gain information about reading habits and special areas of interest (Afflerbach, 2007). Interest assessments are a relevant assessment tool at this stage because learners are beginning to read more widely so teachers should learn what their growing interests are. Though there is scant research literature on interest assessments, Flippo (2003) recommends that they are best developed by the classroom teacher for his or her own learners. However, there are commercial versions such as the Reading Interest Survey available (see Hildebrant, 2001).

Advanced Fluency

Law and Eckles' (1995) final phase in learning to read is the advanced fluency stage. This is when the learner is able to read about unfamiliar topics in a variety of genres. At this juncture, assessment of a variety of genres such as narrative and expository test will be of interest. Caldwell (2002) points out that there are significant differences in the structure of

narrative and expository text with the latter having a much wider variety of structures (e.g., cause and effect, compare and contrast etc.). An assessment technique for evaluating knowledge of text structure that was discussed by Alderson (2000) is to have learners arrange the parts of a text in order. He cautions that these tasks are challenging to create because there could be several unintended orderings of the text.

Secondary School Assessment

Reading Comprehension

At the secondary level, the shift in focus happens whereby students are expected to transition from "learning to read" to "reading to learn." The first corollary of this shift is a stress on the development of reading comprehension and metacognitive strategies. Teachers become increasingly interested in how comprehension strategies can be appropriately assessed. Alderson (2000) suggests two test techniques for assessing reading comprehension. The first is the standard summary. This can be created by having learners choose the best summary, write a partial summary or write their own summary. The strength of this technique is that it allows the learners to express their understanding of the text in their own words. The weakness is that having learners write their own summary makes it into a test of writing as well. The use of summaries for second language assessment is under-researched (Yu, 2007). Nevertheless, the few studies that have been conducted affirm that teaching readers how to summarize aids their ability to effectively summarize (Cohen, 1994) as does providing clearly structured reading texts (Kobayashi, 2002).

A second task for assessing comprehension is information transfer. An example of such a task is having learners identify particular information in the reading passage and transfer it to a

table, chart or some other type of graphic organizer (Alderson, 2000). Research on the helpfulness of information transfer tasks reports that they do have potential for enhancing content area reading comprehension instruction (Mohan, 1986; Jiang & Grabe, 2007). However, there does not appear to be a large research base into their efficacy for first or second language reading assessment.

Metacognitive reading strategies can be divided into those that are used before (e.g., activating background knowledge, predicting), during (monitoring comprehension) and after reading (summarizing). A pre-reading strategy assessment suggested by Valdez Pierce (2000) is an anticipation guide. This is a series of controversial statements about the topic of the reading. Students are asked to discuss their opinion about these statements with their classmates. Anticipation guides can be adapted for assessment by keeping anecdotal records of students' use of pre-reading strategies such as activating prior knowledge or making predictions. A during-reading strategy assessment could take the form of reciprocal teaching (Valdez Pierce, 2000). A review of the research literature on reciprocal teaching supported its use as an assessment strategy with first language learners (Rosenshine & Meister, 1994) but a comparable review of the literature with second language learners does not appear to have been done. First, the teacher explicitly teaches and demonstrates a during-reading strategy such as summarizing or asking clarification questions. Then, in small groups, the students must think aloud to demonstrate the use of reading strategies as they read. The teacher can assess their use of the strategy by using anecdotal records or rating scales (rubrics). An assessment of post-reading strategies suggested by Valdez Pierce (2000) is dialog journals. These are journals where students can provide reactions and ask questions about what they are reading and have it responded to by the teacher.

The teacher can use the form and content of these journals as a source of input for future lessons on areas where students appear to be having difficulty.

Content Area Literacy

A second area of interest in the assessment of secondary-level literacy is content literacy. Content literacy is defined as "the ability to use reading and writing for the acquisition of new content in a given discipline" (McKenna & Robinson, 1990, p. 184). Gunderson (2009) proposes a quick and rough measure of a student's ability to read a particular content area text. He suggests having a student read a short text of approximately 250 words and highlight all the words that he or she does not know. If the learner highlights more than 20 words per 100 then the text is at the frustration level. If fewer than 10 it is at the independent level. Ignorance of any function word (e.g., preposition) would indicate a need for more instruction of basic English. Gunderson (2009) also recommends creating a checklist of all of the skills that students need to read and learn from a content area textbook that they are expected to read in any of their classes. Then, make up an assessment that covers each of the required skills. After that, tell the students that this assessment will help their teacher learn more about what they what they need to know to help them more efficiently read a textbook. Remind them that they will not be formally graded so they can relax. Results of the assessment should give the teacher an overall sense of the skills that students have yet to learn to effectively read their textbook.

Vocabulary Assessment

A crucial aspect of helping students to learn to cope with content area literacy is helping them to comprehend more prevalent technical vocabulary. Despite the continued popularity of traditional approaches to vocabulary assessment around the world, newer trends in vocabulary assessment are beginning to emerge. The traditional view of vocabulary instruction is best

summarized by Read (2007) who notes that “in some respects vocabulary testing is quite a simple activity, a matter of selecting a suitable number of target words and assessing whether each one is known by means of an established test format such as multiple-choice, matching, gap-filling, or some form of translation” (p. 106). Over the past several decades this simple approach to vocabulary assessment has met with increasing criticism from communicative methodologists and test developers. This condemnation is largely due to certain assumptions about vocabulary knowledge that were incorporated into these traditional forms of vocabulary assessment. Namely, that vocabulary had objective meanings that could be measured without any reference to any particular context (Read & Chapelle, 2001).

Critics contend that discrete vocabulary measures needed to be supplemented by interactionalist (i.e., contextualized) measures of vocabulary knowledge. Language assessments needed to be able to determine whether or not language learners can actually use the vocabulary they claim to know in meaningful context that is as authentic as possible (Qian, 2008; Read & Chapelle, (2001). According to Qian (2008), contextualized vocabulary tests differ from traditional vocabulary tests in two main ways. Traditional tests tend to be discrete point and attempt to elicit the meaning of single isolated words that are minimally contextualized. By contrast, interactionalist tests attempt to measure the meaning of words that are placed into an appropriate semantic, grammatical, and communicative context.

Read (2007) discussed several of these newer communicative approaches to vocabulary assessment. The first of these was "yes/no format" vocabulary questions. For this types of assessment, all learners need to do is read a list of words and specify whether or not they know the words by answering yes or no. Some nonwords are included in the list to ensure that learners

are not overestimating the amount of vocabulary that they know. This type of test has found to be useful and economical (Read, 2007). "Word associates format" is another type of assessment that works by using a target word and six or eight other words. Half of the other words are semantically related to or collocate with the target word and half do not. The learner simply has to decide which words relate to the target. These items are designed to assess deep word knowledge. Some research has shown that this type of assessment may be a good alternative to multiple-choice vocabulary tasks (Qian & Schedl, 2004). A third type of vocabulary assessment is the "Vocabulary Knowledge Scale." (Paribakht & Wesche, 1997) This uses a self-report likert-type measure to assess learners' knowledge of target words. This type of assessment can also be combined with other traditional tasks such as providing a synonym.

In a study that compared learners' performance on traditional vocabulary tasks and interactionalist vocabulary tasks, Qian (2008) found that both types of assessment provide similar and statistically significant amounts of prediction for test takers' reading performance but there were particular advantages offered by modern contextualized assessments that traditional decontextualized assessments lacked. One major advantage of contextualized assessments was positive wash back. That is, learners who have to answer questions about vocabulary in context will probably study vocabulary in context for the test.

Computer Assisted Language Assessment for EALs

According to Ockey (2009), computer-based assessment (CBT) is "The use of computers to deliver, score, select items, and report scores of assessments…" (p. 836) Jamison (2005) points out that when CBT's were initially developed they were simply computerized versions of their paper-based counterparts but they have since evolved. Chapelle (2008) observes that the

main reason why CBT was developed was to “improve the efficiency of current testing practice.” (p. 127)

Computer-assisted assessment was not quickly adopted by applied linguists largely because of undeveloped hardware and software and apathy in the large testing programs that had the resources to develop these kinds of tests (Chapelle, 2008). It was not until more recently (mid 1990s) that computer-assisted language tests began to be used for both large and small scale language assessment (Dooey, 2008).

Despite this slow start, the use of computerized forms of assessment has become commonplace. Ockey (2009) notes that “The increasing use of CBT has occurred on numerous assessment fronts, from high-stakes, large-scale commercialize tests to low stakes, small-scale assessments available for free on the Internet" (p. 836). Given the growing prominence of information technology in the modern world, Chapelle and Douglas (2006) suggest that “communicative language ability needs to be conceived in view of the joint role that language and technology play in the process of communication” (p. 108), and we need to expand our view of language ability to see it as “the ability to select and deploy appropriate language through the technologies that are appropriate for a situation” (p. 107).

CBT may allow for positive watch back because learners preparing for these sorts of second language assessments could develop their computer literacy in the process. This is because an increase in the use of CBT ensures that “test takers have needed to reorient their test preparation practices to help them prepare for new test items...” (Chapelle, 2008, p. 127).

Computer Adaptive Testing

Jamison (2005) notes that “in many language testing programs, a “value-added” approach to computerization of the paper-and-pencil format was the introduction of computer-adaptive sections” (p. 230). Computer-adaptive testing (CAT) uses a computer based testing system to administer test questions that are ideally suited to evaluate each test taker’s unique abilities (Embretson & Reise, 2000). Computer adaptive testing tries to imitate a test examiner (Wainer, 1990). Good examiners begin by asking questions that are very easy or difficult. Based on the test taker’s answers, a skilful examiner would then choose questions that are more suited to the level of that particular test taker. This is also what a computer adaptive test attempts to do. The CAT is designed to administer questions that are specifically chosen based on answers previously given by the test taker (Ockey, 2009).

Green (1983) discusses several potential advantages of CAT. One is improved test security. Most test takers would be unable to memorize all of the questions for the test beforehand due to the large number of questions involved. Secondly, test takers are appropriately challenged because test questions are pitched to their ability level. Third, new questions can easily be piloted because CAT tests are discrete point and not integrative.

Despite these advantages, there have also been several problems identified over the years that have prevented CAT from being widely adopted. One of the main problems is the psychometric assumptions that CAT violates. These assumptions both relate to item response theory which serves as the theoretical basis of the CAT (Ockey, 2009). The first assumption is unidimensionality which is the notion that for a CAT test to be valid all of the questions must be measuring the same construct. The problem is that language is not one-dimensional. Thus, it is

very easy to conflate the constructs being measured (Osterlind, 2006). The second assumption of item response theory that is easily violated by CAT tests is local independence. This is the notion that knowledge gained from answering one question cannot be used to answer another question. However, as Ockey (2009) observes, for tests with multiple questions based on the same reading passage, there is a very real danger of this happening.

In addition to violations of theoretical assumptions, there are also practical and logistical problems with CAT. For instance, there is also no commonly agreed-upon algorithm for scoring computer adaptive tests and test results can vary depending on the algorithm that is used (Embretson & Reise, 2000). Secondly, there is the issue of resource consumption. In order for these tests to be effectively developed there needs to be a question bank of thousands of piloted questions that test developers can draw on. As the test continues to be developed, this question bank needs to be continually updated with thousands of new questions. The development and maintenance of such a test question bank is quite resource intensive. The second resource consideration is that developing these sorts of tests takes a great deal of expertise to effectively develop, use, and interpret the test (Ockey, 2009).

Web Based Assessment

There may be answers to the challenges presented by CAT. One possible solution, offered by Winke (2006 cited in Ockey, 2009) is a "semi-adaptive" assessment. An example of a semi-adaptive test is the DIALANG. This assessment has two parts. The first part provides a rough assessment of the test taker’s ability and then the test taker is subsequently provided with suitable tests based on the initial ability assessment (Jamison, 2005). Although the DIALANG does not select the subsequent questions based on current answers, it does provide a more

general form of adaptation that promises to be less resource intensive than current versions of CAT.

DIALANG is also a prototypical example of a web-based test (WBT). These types of tests use a web browser on the Internet to administer a test. This form of assessment has several advantages. First, WBT allows test takers to take assessments from the comfort of their own home anywhere in the world with a computer and access to the Internet. Secondly, keeping assessments in a central server location can allow for improved security. Thirdly, test takers who have experience with using the Internet may actually find the web-based interface to be more recognizable than a traditional computer-based assessment. A final advantage of WBT is that web-based assessment may be more inexpensive than traditional computer-based assessment because the tests could be used on pre-existing browsers instead of needing to develop software from scratch (Ockey, 2009).

Tests like the DIALANG are receiving increasing attention as a means of conducting low stakes assessment. The aim of this sort of assessment is to "use technology to change the dynamic between test takers and tests by providing learners a means for finding out how they are doing, what they need to review, and whether they are justified in their level of confidence about their knowledge" (Chapelle, 2008, p. 129). A meta-analysis of research that explored the comparability of web-based and paper-based tests revealed that there were not significant differences between the two (Westrick & Vispoal, 2009).

Problems with Computer Based Assessment

Although computer-based assessment has a number of strengths which were discussed above, there are also several potential limitations to the use of computer assessments that were

discussed in the research literature. One question about computer based assessment has to do with several threats to validity that it faces. Chapelle and Douglas (2006, p. 41) have summarized threats to the validity of CBT which include:

1. Performance on a computer-delivered test may fail to reflect the same ability as what would be measured by other forms of assessment.
2. The types of items that can be developed in computer-assisted formats are different from those that can be developed with other media.
3. Selection of items to be included on an adaptive test by an algorithm may not result in an appropriate sample of test content and may cause test takers anxiety.

Another challenge presented by computer assessment is the ever-changing technology which necessitates scalable hardware and software as well as ongoing technical skill development for assessment administrators (Chapelle, 2008). This makes CBT quite resource intensive. The significant costs of adopting these new technologies for assessment are often passed on to the test taker through increased testing fees. These increased fees can prevent learners from writing the test which, in some cases, made preclude them from writing the test and gaining access to higher education (Chapelle, 2008).

The problem is that access to the resources required for CBT is often determined by the test administrator or test taker's nationality. As Dooey (2008) observes:

> In general, access to computers is not equally divided among different nationality groups. This means that while access to computers worldwide is increasing, this trend is not as rapid in certain parts of the world. Furthermore, certain language groups seem to be less familiar with computers than others. This highlights the widening gap between the computer 'haves and have nots'". (p. 28)

Test taker access to computers facilities required to do well on these tests is not an uncommon problem either. Taylor et al. (2000) estimate that "one quarter to one half of the

students from most regions of the world will likely need help learning how to use English word-processing programs and the Internet once they arrive at North American colleges and universities" (p. 584).

Dooey (2008) emphasizes the importance that computer based tests must measure the test taker's language ability and not his or her computer skills. However, at this point, "we still do not know with any certainty how computer technology in language tests affects individual test takers who may be advantaged or disadvantaged with respect to computer skills" (Douglas & Hegelheimer, 2008, p. 116). To answer this question, research is being conducted on the comparability between paper-based and computer-based forms of assessment. Paper-based assessments may require different strategies than computer-based assessments and this would affect the validity of the test (Chapelle, 2008).

Sawaki (2001) reviewed the research literature on the comparability between paper and computer based assessments. She concluded that the effects of mode of presentation on test performance are complex, and that the literature suggests that these effects: "may be observed in a change in the nature of a test task, in a decision based on a test score, in test completion time, and in test takers' affect…" (2001, p. 13). In her review of issues in second language assessment and CBT, Dooey (2008) similarly reported that computer-based and pen-and-paper IELTS and TOEFL are comparable. However, "this situation assumes a certain level of computer familiarity... Furthermore, there is some evidence to suggest that different abilities are being measured across the two media, even where the scores obtained could be considered comparable" (p. 32-33).

Dooey (2008) also points out the risks of using technology because there is always the possibility that it could fail. She suggests that “pen-and-paper test versions should always be an option, particularly in high-stakes testing. This not only provides alternatives for those administering the test, it also offers test-takers the choice in how they take the test” (Dooey, 2008, p. 33).

Chapelle (2008) raises some important questions about the future role of applied linguists in the development of computer-based second language assessments. She points out that most students of second-language assessment do not have a strong understanding of computer assessment in language learning. She adds that "if graduate students are to dig into the language testing issues, they need to be able to create and experiment with computer-based tests" (Chapelle, 2008, p. 132). Therefore, they need to have a sound understanding of both the technology and the research-proven principles of second language assessment.

Criticisms of Societal Impact of Second Language Assessments

There are two compelling criticisms regarding the design and use of second language tests. One objection relates to the power that tests have over students’ lives. Shohamy (2000) points out that “test results have detrimental effects for test takers since such uses can create winners and losers, successes and failures, rejections and acceptances” (p. 15). She discusses some of the ways that second language assessments are misused to the detriment of learners. For instance, test results often decide whether or not students are eligible to enter a program and how they will be placed if they do. Likewise, tests are often used to judge whether or not the learner is qualified to graduate from the program. Tests are often the gatekeepers to the best universities and occupational opportunities. House (1998) contends that in this way tests serve as a tool by

which powerful elite groups in society maintain control through perpetuating class differences by using tests to keep “undesirables” in their place at the bottom of the hierarchy.

One means by which tests can systematically discriminate against particular groups is through test bias in methods and materials. There are several ways that tests can be biased. One is when the content or concepts being tested are those with which only the dominant group is familiar. Another is the use of language in the test that might be unknown to non-native speakers. Even different language groups may be affected differently by such bias. For instance, Chen and Henning (1985) found that speakers of Asian languages may be disadvantaged in comparison to their European peers whose languages share many more cognates with English. Tests may also be biased in their methods. For instance, use of the western method of the teacher asking display questions may be alien to some cultures. Additionally, a test is biased when it is normed on the dominant group but used on minority groups who were not accounted for in the norming process. Test items that are culturally or linguistically biased against minority groups could inadvertently (or purposely) be in this type of test (Garcia & Pearson, 1994).

According to Garcia and Pearson (1994) and Helms (1992) there are several things that teachers can do to minimize test bias. The first is to use many different methods of assessment. That will allow students more opportunities to demonstrate their ability through a mode that may be better suited to their culturally appropriate ways of displaying knowledge. Secondly, test users need to learn more about what sorts of knowledge and strengths learners from other cultures bring and try to build on those in assessment. Finally, test users must try to find out why learners make the mistakes they do by having them explain their rationale for the answers that they gave on a test question. This may expose any potentially biased assumptions of the test.

Conclusion

The preceding review of the research and professional literature on second language literacy has shown that there is a wide variety of second language literacy assessment tools available which include both traditional and alternative forms of assessment. The literature reviewed here has also demonstrated that there are a wide variety of assessment options available to fit the unique needs of EAL learners at all grade levels. Hopefully, it has also allowed for some reflection on the serious consequences that assessment has on learners' lives and the burden that brings.

Research continues to accumulate for the effectiveness of newer more authentic and communicative forms of assessment but there is still a need for research in second language literacy assessment. In particular, there is a dearth of research into the efficacy of newer adaptations of the cloze method with second language learners or forms of informal classroom assessment such as using graphic organizers and communicative vocabulary assessment tasks.

This review has explored many of the traditional and progressive options that school personnel have at their disposal to diagnose and remediate the literacy needs of EAL learners. It is hoped that these resources can assist educators in the lower mainland in their continued efforts to educate immigrant learners.

References

Abraham,R. G. & Chapelle,C.A. (1992). The meaning of cloze test scores: An item difficulty perspective. *The Modern Language Journal, 76,* 468-479.

Aebersold, J. A. & Field, M. L. (1997). *From reader to reading teacher: Issues and strategies for second language classroom*. Cambridge: Cambridge University Press.

Afflerbach, P. (2007). *Understanding and using reading assessment, K-12*. Newark, DE: International Reading Association.

Alderson**,** J.C. (2000**).** *Assessing reading.* New York: Cambridge University Press

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge: Cambridge University Press.

Babaii, E. & Ansary, H. (2001). The C-test: A valid operationalization of reduced redundancy principle? *System*, *29*, 209-219.

Bachman, L.F. (1982). The trait structure of cloze test scores. *TESOL Quarterly*, *16*, 61–70.

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

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

Bailey, K. M. (1998). *Learning about language assessment: Dilemmas, decisions, and directions.* Toronto: Heinle & Heinle.

Baker, D. (1989). *Language testing: A critical survey and practical guide.* New York: Hodder & Stoughton.

Bialystok, E., Shenfield, T., & Codd, J. (2000). Languages, scripts, and the environment: Factors in developing concepts of print. *Developmental Psychology*, *36*, 66-76.

Blanche, P. & Merino, B. (1989). Self-assessment of foreign language skills: implications for teachers and researchers. *Language Learning*, *39*, 313–40.

British Columbia Teachers' Federation Information Services & British Columbia Teachers' Federation Research Department (2009). *BC education fact sheet.* British Columbia Retrieved May 15, 2010, from http://bctf.ca/uploadedFiles/public/Publications/2009EdFactSheet.pdf

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment.* New York: McGraw-Hill.

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

Caldwell, A. S. (2002). *Reading assessment: A primer for teachers and tutors.* New York: Guilford.

Camp, R. (1993). The place of portfolios in our changing views of writing assessment. In R. E. Bennett, & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 183–212). Hillsdale, NJ: Lawrence Erlbaum.

Carrol, J. B. (1986). LT+25, and beyond? Comments. *Language Testing, 3*, 123-129.

CBC. (2010, 10 March). *Minorities to rise significantly by 2031: StatsCan*, from http://www.cbc.ca/canada/story/2010/03/09/statscan-minority.html

Centurion, C. E. (1985). *The use of the informal reading inventories and miscue analysis to evaluate Oral reading, reading comprehension and language inference/transfer in English as second language students. Unpublished doctoral dissertation*, Texas A&I University, United States, Texas. (Publication No. AAT 8527088) Retrieved March 30, 2010, from Dissertations & Theses database.

Chapelle, C. A. (2008). Utilizing technology in language assessment. In E. Shohamy & N. H. Hornberger (Eds.). *Encyclopaedia of language and education*, 2nd Edition, Volume 7: language testing and assessment.

Chapelle, C. A. & Douglas, D. (2006). *Assessing language to computer technology.* Cambridge: Cambridge University press.

Chavez-Oller, M.A., Chihara, T., Weaver, K.A. & Oller, J.W. (1994). When are cloze items sensitive to constraints across sentences? In Oller, J.W. Jr. & Jonz, J., (Eds.), *Cloze and coherence* (pp. 229–45). London: Associated University Press.

Chen, Z. & Henning, G. (1989). Linguistic and cultural bias in language proficiency tests. *Language Testing, 2*, 155-163.

Cherry, R,, & Meyer, P. (1993). Reliability issues in holistic assessment. In M. Williamson & B, Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 109-141). Cresskill. NJ: Hampton Press

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

Dooey, P. (2008). Language testing and technology: Problems of transition to a new era, *ReCALL, 20,* 21-34.

Dornyei, Z. & Katona, L. (1992). Validation of C-test among Hungarian EFL learners, *Language Testing, 9,* 187-206.

Douglas, D., & Hegelheimer, V. (2007). Assessing language using computer technology. *Annual Review of Applied Linguistics, 27,* 115–132.

East, M. & Young, D. (2007). Scoring L2 writing samples: Exploring the relative effectiveness of two different diagnostic methods. *New Zealand Studies in Applied Linguistics, B, 1-21*

Embretson, S., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Espin, C. A. & Froegen, A. (1996). Validity of general outcome measures for predicting secondary students' performance on content-area tasks. *Exceptional Children, 62,* 497-514.

Flippo, R. F. (2003). *Assessing readers: Qualitative diagnosis and instruction.* Portsmouth, NH: Heinemann.

Freedman, S., & Sperling, M. (1985). Written language acquisition: The role of response and the writing conference. In S. Freedman (Ed.), *Acquisition of written language: Response and revision* (pp. 106-130). Norwood, NJ: Ablex.

Freeman, M. (1995). Peer assessment by groups of group work. *Assessment and Evaluation in Higher Education,* 20, 289–99.

Garcia, G. E. & Pearson, P. D. (1994). Assessment and diversity. In L. Darling-Hammond (Ed.), *Review of research in education* (pp. 337-392). Washington: American Educational Research Association.

Genessee, F., & Upshur, J. (1996). *Classroom-based evaluation in second language education.* Cambridge: Cambridge University Press.

Goldstein, L. & Conrad, S. M. (1990). Student input and negotiation of meaning in ESL writing conferences. *TESOL Quarterly, 24,* 443-460.

Green, B. (1983). The promise of tailored tests. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement* (pp. 69–80). Hillsdale, NJ: Erlbaum.

Gunderson, L. (2007). *English-only instruction and immigrant students in secondary schools: A critical examination.* Mahwah, NJ: Lawrence Erlbaum Associates.

Gunderson, L. (2009). *ESL (ELL) literacy instruction: A guidebook to theory and practice (2nd ed.).* New York: Routledge.

Helms, J.E. (1992). Why is there no study of cultural equivalence in standardized cognitive ability testing? *American Psychologist, 47*, 1083-1101.

Hildebrant, D. (2001). But there's nothing good to read (in the library media center). *Media Spectrum: The Journal for Library Media Specialists in Michigan, 28,* 34-37.

House, E. (1998). *Schools for sale*. New York: Teacher College Press.

Hughes, A. (1989). *Testing for language teachers.* Cambridge: Cambridge University Press.

Hurley, S. R. & Tinajero, J. V. (Eds.). *Literacy assessment of second language learners* (pp. 64-83). Needham Heights, MS: Allyn & Bacon.

Jamison, J. (2005). Trends in computer-based second-language assessment. Annual Review of Applied Linguistics, 25, 228-242.

Jiang, X. & Grabe, W. (2007). Graphic organizers in reading instruction: Research findings and issues. *Reading in a Foreign Language, 19,* 34–55.

Jonz, J. (1990). Another turn in the conversation: what does cloze measure? *TESOL Quarterly, 24*, 61–83.

Kalia, V. (2007). Assessing the role of book reading practices in Indian bilingual children's English language and literacy development. *Early Childhood Education Journal, 35,* 149-153.

Kalia, V. & Reese, E. (2009). Relations between Indian children's home literacy environment and their English oral language and literacy skills. *Scientific Studies of Reading, 13,* 122-145.

Kern, R. (2000). *Literacy and language teaching.* Oxford: Oxford University Press.

Klingner, J. K., Artiles, A. J., & Barletta, L. M. (2006). English language learners who struggle with reading: Language acquisition or LD? *Journal of Learning Disabilities, 39,* 108-128.

Kobayashi, M. (2002). Method effects on reading comprehension test performance: text organization and response format. *Language Testing, 19*, 193-220.

Kwan, K. & Leung, R. (1996). Tutor versus peer group assessment of student performance in a stimulation training exercise. *Assessment and Evaluation in Higher Education, 21*, 239–49.

Law, B. & Eckes, M. (1995). *Assessment and ESL: On the yellow big road to the withered of Oz.* Manitoba: Peguis.

Lee, Y.J. (2006). The process-oriented ESL writing assessment: Promises and challenges. *Journal of Second Language Writing, 15*, 307–330.

Markham, P.L. (1985). The rational deletion cloze and global comprehension in German. *Language learning, 35*, 423-430.

McBeath, A. (1989). C-tests in English: Pushed beyond the original concept? *RELC Journal, 20*, 36-41.

McKenna, M. C. (1983). Informal reading inventories: A review of the issues. *The Reading Teacher, 36*, 670-679.

McKenna, M.C. & Layton, K. (1990). Concurrent validity of cloze as a measure of intersentential comprehension. *Journal of Educational Psychology, 82*, 372–77

McKenna, M. C. & Robinson, R. D. (1990). Content literacy: A definition and implications. *Journal of Reading, 34*, 184-186.

McNamara, T. (1996). *Measuring second language performance.* New York: Longman

Mohan, B. A. (1986). *Language and content*. Reading, MA: Addison-Wesley.

Neill, D. M. & Medina, N. (1989). Standardized testing: Harmful to educational health. *Phi Delta Kappa International, 70*, 688-697.

Newman, C., Smolen L., Lee D. J., Aron, V. (1996, February). *Student maintained portfolios and peer mentoring as a means of empowering and motivating students: Unexpected outcomes.* Paper presented at the Eastern Educational research Association Annual Meeting, Boston MA.

Ockey, G. J. (2009). Developments and challenges in the use of computer-based testing for assessing second language ability, *The Modern Language Journal, 93*, 836-847.

Oikonomidoy, E. (2007). I see myself as a different person who [has] acquired a lot …': Somali female students' journeys to belonging. *Intercultural Education, 18*, 15–27.

Oller, J. W. (1979). *Language tests at school.* London: Longman.

Oller, J. W. & Jonz, J. (Eds.).(1994). *Cloze and coherence*. Cranbury, NJ: Bucknes University Press.

Orsmond, P., Merry, S. & Reiling, K. (1997). A study in self assessment: tutor and students' perceptions of performance criteria. *Assessment & Evaluation in Higher Education* 22, 357–67.

Osterlind, J. (2006). *Modern measurement: Theory, principles, and applications of mental appraisal* . Upper Saddle River, NJ: Pearson Merrill Prentice Hall.

Paribakht, T. S., & Wesche, M. (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary development. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 174–200). New York: Cambridge University Press.

Propst, I. & Baldauf, R. B. (1979). Use matching cloze tests for elementary ESL students. *The Reading Teacher, 32*, 683-690.

Qian, D. D. (2008). From single words to passages: Contextual effects on predictive power of vocabulary measures for assessing reading performance. *Language Assessment Quarterly, 5*, 1–19.

Qian, D.D., & Schedl, M. (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing, 2*1, 28-52

Radwanski, G. (1987). *Ontario study of the relevance of education and the issue of dropouts.* Toronto: Ontario Ministry of Education.

Read, J. (2007). Second language vocabulary assessment: Current practices and new directions. *International Journal of English Studies.* 7, 105-125.

Read, J. & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing, 18*, 1–32.

Rosenshine & Meister, C. (1994). Reciprocal teaching: A review of the research. *Review of Educational Research, 64*, 479-530.

Ross, S. (1998). Self-assessment in second language testing: a meta-analysis and analysis of experiential factors. *Language Testing, 15*, 1–20.

Sawaki, Y. (2001). Comparability of conventional and computerized tests of reading in a second language. *Language Learning & Technology, 5*, 38-59.

Shanahan, T., Kamil, M.L. & Tobin, A.W. (1982). Cloze as a measure of intersentential comprehension. *Reading Research Quarterly, 17,* 229–55.

Shohamy, E. (2000). *The power of tests: A critical perspective on the uses of language tests*. Essex: Pearson.

Slater, S. J. (1980). Introduction to performance testing. In Spirer, J. E. (ed.). *Performance testing: Issues facing vocational education* (pp. 3-17). Columbus, OH: National Center for Research in Vocational Education.

Song, B. & August, B. (2002). Using portfolios to assess the writing of ESL students: A powerful alternative? *Journal of Second Language Writing, 11,* 49-72

Specter, J. (2005). How reliable are informal reading inventories? *Psychology in the Schools*, 42, 593-603.

Steinman, L. (2002). A touch of...class! *The Canadian Modern Language Review, 59*, 921-301.

Taylor, C., Jamieson, J. and Eignor, D. (2000) Trends in computer use among international students. *TESOL Quarterly, 34*, 575–85.

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68, 249-276.*

Valdez Pierce, L. (2000). Assessment of reading comprehension strategies for intermediate bilingual learners. In Hurley, S. R. & J. V. Tinajero (Eds.). *Literacy assessment of second language learners* (pp. 64-83). Needham Heights, MS: Allyn & Bacon.

Wainer, H. (1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Erlbaum.

Watt, D. & Roessingh, H. (2001). The dynamics of ESL dropout: Plus ca change... *Canadian modern language review*. 58, 203-222.

Weigle, S. C, (2002). *Assessing writing.* Cambridge: Cambridge University Press.

Weir, C. (2005), *Language testing and validation: An evidence-based approach.* Basingstoke: Paigrave Macmillan.

Wesche, M. B. (1987). Second language performance testing: the Ontario Test of ESL as an example. *Language Testing, 4*, 28-47.

Westrick P. A. & Vispoal, W. P. (2009, August) *Meta-analysis of web-based and paper-based measures.* Paper presented at the 117th Annual Convention of the American Psychological Association, Toronto, ON.

Worthen, B (1993). Critical issues that will determine the future of alternative assessment, *Phi Delta Kappa International, 74*, 450-454.

Yamashita, J. (2003). Processes of taking a gap-filling test: comparison of skilled and less skilled EFL readers. *Language Testing, 20*, 267-293.

Yu, G. (2007). Students' voices in the evaluation of their written summaries: Empowerment and democracy for test takers? *Language Testing*, 24, 539–572.

Zamel, V. (1985). Responding to student writing. *TESOL Quarterly, 19*, 79-97.