Considerations on the Effective Design and Application of Language Tests

Michael James Davies

Testing is an integral part of any language program and when used effectively serves several important functions. As a result, teachers test their students all the time, either formally or informally, in order to make certain evaluations. These can include, for example, measuring the progress that students are making during a particular course; in other words, to determine whether students are reaching goals laid out in the initial course specifications. Testing can also give the teacher valuable feedback as to the effectiveness of their teaching methods as well as the materials they are using such as textbooks and audiovisual aids. Indeed, Bachman considers "accountability and feedback as essential mechanisms for the continued effectiveness of any educational program" (Bachman, 1990, p. 55). There is therefore an underlying assumption that through testing the educational content of any course can be improved and hence the learning experience of students enhanced. A discussion of the various types of testing is beyond the scope of this paper and we will therefore treat testing in more general terms and as essentially a tool for evaluation. This paper will deal with the fundamental considerations that have to go into any design of language tests. These include such factors as reliability, validity, authenticity and the promotion of positive washback. Therefore, an explanation of these will follow. Following on from this, a real test situation will be critically discussed in terms of two possible marking schemes in order to exemplify the effective use of the aforementioned

considerations. By doing this, the reader will be able to understand how they can be practically put to use in the design of language tests.

Reliability, in the words of Henning, "has been shown to be another word for consistency of measurement" (Henning, 1987, p. 75). Indeed, as Brown states, "a reliable test is consistent and dependable" (Brown, 2007, p. 447). It is an unfortunate truth that a student taking a test on a certain day would score differently if he or she had hypothetically taken the test the day before. This is unavoidable but needn't lead us to the assumption that tests are thus inherently unreliable and bereft of meaning. A lot can be done to increase the reliability of tests in order to minimize any fluctuations in scoring. In fact, it is possible to measure the reliability coefficient of a test that can then be used to elucidate standard deviations. The use of the latter can result in much fairer evaluations of students especially if important decisions rest on such results. According to Brown, there are four main sources of potential unreliability and these are:

- The test itself
- The administration of the test
- The test-taker
- The scoring of the test

(Brown, 2007, p. 447)

As for the test itself, there are a number of ways in which we can endeavor to increase reliability. Firstly, the test should have enough items: generally speaking, the more items included in the test, the greater the reliability up to a point of asymptote, after which the graph plateaus and no further increase in reliability can be observed. This will also, in the words of Henning, provide "greater person separability" and thus "less likelihood that examinees would change rank order on repeated administra-

tions of the test" (Henning, 1987, p. 78). It is also important that tests are of suitable difficulty, i.e., not too easy or not too difficult where students are bunched together at one end of the scoring continuum or the other. Such tests are unreliable because they make the discrimination of ability between the test-takers almost meaningless. Another consideration when making tests is to give the students 'fresh starts'. This means that items should as much as possible be independent of one another where the answer to one item should not have a direct bearing on the test-takers' ability to answer the following question. This can impede reliability, as can giving test-takers too much freedom in expressing their answers. This is particularly true in writing tests where the narrower the field of questioning, the more reliable the test-taking is sure to be. Furthermore, ambiguity in questioning should be avoided and great care expended in ensuring that answers other than the correct one are not acceptable. This is especially the case with regards to multiple choice items where the distracters should be unambiguously unacceptable. These are just a few of the ways in which the test itself can be designed in such a way as to promote reliability. Once a test has been designed, it is always a good idea for students to familiarize themselves with the format of any test as, according to Hughes, "if any aspect of the test is unfamiliar to candidates, they are likely to perform less well than they would do otherwise" (Hughes, 2007, p. 47). Another source of test unreliability is the inconsistency of test administration. Such fluctuations can exist among institutions administering the same test, or even within them. Hence it is imperative that prior to the giving of the test, certain ground rules have been laid and the test administrators have a clear, unambiguous and unified code by which to administer the test. Ideally, test administrators should be trained beforehand, but failing this, according to Henning, "at least written guidelines for test administration should be supplied to all administra-

tors" (Henning, 1987, p. 77). Furthermore, differences in the environment can introduce unreliability into a test: the lighting in a room, the audibility of the CD in a listening test etc. All of this has to be carefully considered prior to the test. According to Hughes, the test-taker can also contribute unwittingly to increasing test unreliability by factors such as sickness, fatigue or emotional disturbance, which, in the words of Henning would cause their score to "temporarily deviate from his or her true score, or that score which reflects his or her actual ability" (Henning, 1987, p. 76). Henning goes on to recommend providing a test environment that is physically and emotionally comfortable for the candidate. Finally, to bring this discussion of test reliability to a conclusion, it is important to raise the issue of scorer reliability. This perhaps is one area where the greatest level of unreliability can be introduced into the test. This is particularly the case where a degree of subjectivity is allowed in the scoring (e.g. a writing test). Variations among scorers are the most obvious danger here, so-called inter-rater error variance. Hughes states that the training of scorers is of the utmost importance in such a case and suggests that the scoring of compositions "should not be assigned to anyone who has not learned to score accurately compositions from past administrations" (Hughes, 2007, p. 49). In addition to this, error may be reduced "by employing detailed rating schedules for the independent use of all judges" (Henning, 1987, p. 77). There is also the consideration of intra-rater error variance through such factors as fatigue, lack of experience, even unconscious favoritism towards candidates (if the name of the test-taker appears on the test). Indeed, the idiosyncrasies of individuals will always have an effect on the reliability of scores and this is why Hughes suggests that multiple, independent scoring is preferable in that any discrepancies in the scores can be fully investigated. To sum up, this paragraph has highlighted some of the more salient threats to test reliabil-

ity and how they can be adequately counteracted. Reliability in a test is of paramount importance and must be considered carefully during the design stage. It is also very important that not only the test itself but its administration and subsequent scoring should be well thought out and thoroughly applied in order to maximize reliability.

Another major consideration when designing a test is to ensure that it has validity. In other words, confirm that the test is actually testing what it is supposed to test. This may seem so obvious as to be hardly worth mentioning, but it is indeed an important issue in good test design. For example, if a listening test is administered to candidates and their written answers are penalized for poor spelling or punctuation, then this is clearly not a very valid test; after all, we are purporting to test their listening ability and not their writing ability. In this paragraph, the major types of validity will be discussed as well as the pitfalls that need to be avoided in order to maintain validity. In the words of Hughes, "we create language tests in order to measure such essentially theoretical constructs as 'reading ability', 'fluency in speaking', 'control of grammar' and so on (Hughes, 2007, p. 26). Therefore, if a test attempts to test these constructs it is deemed to have construct validity. However, this in itself is insufficient and is predicated on such subordinate measurements as content validity and face validity. When a test endeavors to mirror as much as possible the course specifications that the candidates have taken, it is said to have content validity. Of course, for the practical reasons of time and expense, it is not often possible to test every specification, but as long as a representative sample appears in the test, then this can be sufficient. Great care though has to be taken when choosing a representative sample of items lest it results in harmful washback (discussed later) by ignoring large areas of the course specifications. As Hughes also points out, "too

often the content of tests is determined by what is easy to test rather than what is important to test" (Hughes, 2007, p. 27). Thus it is the responsibility of the test maker to shun expediency and ensure a true, representative sample of items appears in the test. Further to content validity is the matter of criterion validity; in other words, the question of whether the test-taker's test score correlates with an independent assessment criterion. This could be as simple as comparing it to ongoing assessment marks kept by a teacher, where those assessments have been competently scored. Where a student is required to take a test in order to be placed in one of various classes depending on ability, the criterion validity can be elucidated by seeing how well the students perform once they have been placed. If a large number of the students have been clearly placed in inappropriate classes, then the initial placement test can be said to have poor criterion validity. This form of validity is actually empirical in that data can be obtained in order to derive a validity coefficient. This coefficient can then be used to discover if there is a high or low level of agreement with the independent assessment criterion. Other forms of validity which may not be empirical as such, but nevertheless can be highly influential, include 'face validity' which is explained by Brown in terms of the question, "Does the test, on the face of it, appear from the learner's perspective to test what it is designed to test?" (Brown, 2007, p. 449). This is hardly a scientific concept, yet can prove highly influential in whether a test is accepted or not by teachers and students alike. Tests which purport to be testing certain abilities by notably indirect methods are particularly susceptible to be rejected on their face value alone. Another nonempirical validity is 'response validity' which according to Henning, "is intended to describe the extent to which examinees responded in the manner expected by the test developers" (Henning, 1987, p. 96). Candidates who adopt a poor attitude toward the test and fail to exert

their best efforts are going to have a marked influence on the validity of the test. These then are some of the major forms of validity that need to be addressed in good test design. The pitfalls that one can avoid when attempting to adhere to these notions are numerous and a few have already been introduced. Inappropriate selection of content is one such obvious area, and happens, according to Henning, "when items do not match the objectives or the content of instruction" (Henning, 1987, p. 91). For example, asking students, in an achievement test, to answer questions on areas that have not been covered during the course is hardly going to enhance the validity of the test. Neither is the misapplication of tests in which highly valid tests may be applied in inappropriate circumstances. Another possible threat to validity is, as Henning describes it, "inappropriate referent or norming population" (Henning, 1987, p. 92). Standardized tests are often developed by using specific subjects, perhaps chosen for their language or cultural background. Tests that have been developed by using one specific population may thus lack validity if administered to another. Thus, it is important that the norming population that the test is designed for is consistent with that it is administered to. As Henning rightly states, "it follows that careful consideration must always be made of the referent or norming population when selecting standardized tests for any given purpose (Henning, 1987, p. 92). This paragraph has therefore expanded on the crucial area of validity in language testing. It has covered some of the more salient forms of validity, such as construct validity, content validity and criterion validity. It then addressed possible problems that may arise in test design that could threaten this test validity.

Further to the issues of reliability and validity when designing a test are those of authenticity and washback. What is meant by authenticity is the degree to which the test is applicable to the real world and thus tests

the ability of the candidate to apply his or her skills and knowledge to practical, meaningful ends. While authenticity is a rather nebulous concept to define, in recent years we have witnessed a move in the classroom towards a more content-based communicative approach that would seem to represent such a definition. Such testing may be viewed by teacher and student alike as more interactive and stimulating and thus more likely to foster communicative competence. Compare this to testing a few years ago that too often, in the words of Brown, used "unconnected, contrived, boring items" to test a "grammatical form or lexical item" (Brown, 2007, p. 451). Today, the move towards more real-world, 'authentic', material in the classroom has marked a real shift in the manner in which language is now being taught and tested. Tests such as the iBT (Internet-based) TOEFL now use a more integrated format that reflects the kinds of situations that students will encounter in a foreign university classroom. For example, the ability to derive the main points from a lecture or a short reading; using and distilling such information to form a short speech or writing passage; and being able to express one's opinions on real-world matters. Such trends in teaching and testing are to be applauded but the caveat is that so often, practical considerations are often an obstacle to the making of such tests. The use of tests that integrate the four macroskills can often lead to questions of reliability. In addition, it can also lead to difficulties in evaluation; a mistake in an oral test may be due to either a listening error or a speaking error, one can never be sure. However, despite the practical difficulties that undoubtedly exist in the design of such tests it is still a goal worth working towards. Tests that require candidates to confront real-world tasks are more likely to promote the ultimate goal of communicative competence. This can be achieved by the concept of washback, where the teaching of a program is influenced by the content of the test itself. Far from being a negative idea, washback can be

very much a force for good as the rest of this paragraph will argue. The importance of washback has been recognized in recent years and a great deal of research has since been carried out on it. As Hughes states, washback "is now seen as a part of the impact a test may have on learners and teachers, on educational systems in general, and on society at large" (Hughes, 2007, p. 53). Indeed, it is now a crucial consideration in the design of tests as its influence on what is taught to students during the program prior to the test cannot be underestimated. Tests that habitually include only items that measure understanding of a small proportion of the test specifications may not be encouraging positive washback, especially if the questions are similar year in, year out. Teachers may choose to ignore chunks of the test specifications preferring only to focus on those areas that frequently appear in the final test. Thus, a good test will include items that have been chosen widely and unpredictably from the test specifications. In addition, tests that are well designed will also encourage the explicit teaching of skills such as speaking or writing. For example, a test that purports to measure a candidate's ability to write in English should thus require them to actually write in English. This may sound absurdly obvious, but it is surprising that due to such practical considerations as time and difficulties with objectivity, tests may not actually directly measure the skill they were designed to test. Thus, negative washback occurs in that teachers will teach to the test and discourage students from actually practicing the skill they are trying to improve. Another aspect of washback that must be mentioned is the potentially good influence of thorough feedback after the test. So often, teachers will hand back tests or homework assignments in which they have merely added a grade or a percentage. Such feedback is unlikely to encourage improvement in students who have no idea why they were marked in a particular way. Brown suggests giving praise for strengths as well as constructive criti-

cism where necessary. According to him, teachers should "take some time to make the test performance an intrinsically motivating experience through which a student will feel a sense of accomplishment and challenge" (Brown, 2007, p. 452).

The previous paragraphs in this paper have endeavored to highlight, in theory, what is required in the design and application of an effective language test. It has, in particular, dwelt on four factors, namely: reliability, validity, authenticity and washback. The following paragraphs will now place this theory into practice by critically looking at a test in terms of the previously mentioned factors. The critique will be organized into three main sections. These will comprise critiques of: the task itself; the first marking approach; and finally, the second marking approach. These sections will be discussed with respect to their reliability, validity, authenticity and potential for achieving effective washback. It is the argument of this report that while the writing task itself may have certain flaws that will be discussed forthwith, it is at least encouraging direct testing. Furthermore, the second marking approach is by far the better of the two for reasons which will also be laid out. Let us first look at the task:

> The following writing task was used to assess the basic writing skills of 90 teenage students.
>
> Think of a time when you were traveling somewhere and the journey was very long. Discuss your experience.
>
> Marking approach 1
> Scripts were divided equally between all teachers working with these students to mark in their non-teaching time. They were asked

to provide a score for each of the following language features and then allocate a total score out of 25 and then covert that score to a percentage.

(a) spelling

(b) punctuation

(c) grammar

(d) control

(e) argument

Marking Approach 2

All scripts were marked by two teachers of English. They considered the following language features: spelling, punctuation, grammar, vocabulary, cohesion, sentence structure and overall impression. On the basis of quickly reading each script they grouped them on the basis of their overall impression. To mark the various language features they applied a range of criteria which had previously been developed to link with identified levels of proficiency. Thus, multiple scoring was in use and markers were provided with descriptive criteria to guide the allocation of scores. For example, when marking for cohesion they used criteria such as that noted in O'Neill and Gish (2008, p. 247) :

*Score 2 when there is use of complex sentences, lack of repetitious use of 'and' and 'then', use of more sophisticated cohesive ties such as however, although, in fact, first, secondly, usually, after, before, as soon as, until, while, eventually, during, meanwhile, thus, consequently and therefore, and there is evidence of use of some variety of cohesive ties and the piece of writing conveys a sense of completeness.

*Score 1 when there is either use of predominantly simple sentences

> or simple sentences and some complex sentences which provide evidence of connecting ideas through the use of basic ties such as and, then, so, but, also, next, when, because and suddenly.
> *Score 0 when there is little or no evidence of use of cohesive ties and/or connecting of ideas, lack of sense of wholeness, illegible responses or one which is irrelevant to the set topic.
>
> When all scripts were marked they compared their scoring and if there were differences they reviewed the script in relation to the marking criteria and arrived at a mutually agreed upon score.
>
> (Mangubhai, 2008)

Let us first of all consider the task that is required of the teenage test-takers to carry out. It is impossible to be absolutely certain but I would imagine that the test designer's intention here is to elicit details of a long journey, be it by plane, train, or car, perhaps in a foreign location, and discover how the candidate felt at that time. This may include feelings of fascination with new sights and scenery, interesting encounters with new people, perhaps occasional boredom with the monotony of a long journey. The scope here for descriptive and expressive writing could indeed be wide. However, it may be that the candidate simply writes about his or her journey to the exam center that morning, or his or her day to day commute to school. Thus, a major criticism of this task is that it may potentially produce answers unanticipated by the test designer. Furthermore, the sheer breadth of possibilities here may well, as stated by Hughes, "have a depressing effect on the reliability of the test" (Hughes, 2007, p. 45). If the test designer had taken more control over the task by adding specifics, then the freedom of the candidate would have been more

restricted and the reliability of the test improved. Moreover, it would have resulted in a test that, in the words of Hughes, is "likely to be a much more reliable indicator of writing ability" (Hughes, 2007, p. 46). Perhaps this could be restricted by replacing *travelling somewhere with taking a trip during the school holidays*. Many teenagers are unlikely to have travelled so widely or exotically and the proximity of school holidays in their memories may make it easier to recall appropriate material. There may be a question of authenticity here, in that we are assuming that such youthful candidates have experienced a long journey and will find the task relevant to their experience.. It is the relevance of the task to test-takers, as stated by Bachman and Palmer, "that…helps promote a positive affective response to the test task and can thus help test-takers perform at their best" (Bachman and Palmer, 1996, p. 24). It is crucial that a test is considered authentic in that it acts as a bridge between what is being tested and the TLU (target language use) domain in which test-takers will be using the language. According to Bachman and Palmer "it is this correspondence that is at the heart of authenticity" (Bachman and Palmer, 1996, p. 23) and needs to be addressed carefully. Thus, in its present format the task may be open to questions of authenticity. Thus, if the task is to be used, then I would suggest that it could be improved by steering the students towards tasks that may be perceived more relevant for them as it is related to "language use in the TLU domain, or to other similar non-test language domains (Bachman and Palmer, 1996, p. 24). For example, the test-taker may be asked more specifically about who they travelled with, what they saw on the journey, who they met, what they ate, where they slept and so on. This would all help to alleviate some of the ambiguity of the task. Having said all that, the task has merit in that it is an example of direct testing and can claim content validity. In the words of Brown, this means that "it requires the test-taker to perform the

behavior that is being measured" (Brown, 2007, p. 449). Furthermore, in order to demonstrate their writing skills, the students are actually having to produce a piece of writing which should have a beneficial washback effect; any course taken to prepare students for such a test would necessarily have to provide plenty of opportunity for writing compositions.

Let us now move on to the first marking approach. This approach would appear to contain a number of flaws. The reliability of such a system has to be brought into question, not least because the scripts are only marked one time. It is also unclear as to whether all the teachers involved will actually be teachers of English; it may be the case that teachers of other subjects are asked to help with the scoring. As the scripts are divided equally among those teachers who work with the students, it is clear that some teachers will end up marking their own students' work as well as those students they don't teach. As Henning points out, "if the rater knows the examinee and the examinee's name appears on the paper, personality factors may influence the scoring process" (Henning, 1987, p. 76). Brown also talks about the possible "bias toward particular 'good' and 'bad' students" (Brown, 2004, p. 21). Even in the case where names are replaced by numbers, there is still the possibility of recognition from handwriting. Hence as Henning calls it, "Intra-Rater Error Variance" is a very real consideration when taking scoring into account (Henning, 1987, p. 76). Such variance can also be exacerbated by fatigue. The teachers will be marking these exams in their non-teaching time as opposed to a time when they may not be teaching at all, or have a lighter load (e.g. school holidays). As a result, "the rater himself or herself is liable to become less accurate with fatigue" (Henning, 1987, p. 76). There is also the problem of error between scorers, in particular in this case where the scoring key is vague to start with. The five criteria on

which the teachers are asked to rate the scripts include such factors as *control and argument*. It is difficult to see where argument plays a role in a task which is simply asking the test-taker to describe his or her experience of a journey. Even if it were the case that the task required an argument, it would then cease to be wholly a writing test. With regards to asking students about their opinions or general knowledge in writing tasks, Hughes states that "for the sake of validity, we should not set tasks which measure these abilities" (Hughes, 2007, p. 90). Furthermore, the actual meaning of 'control' is omitted here, leaving much open to the interpretation of the individual teacher. The first three criteria, those of spelling, punctuation and grammar would seem to lend themselves more to the assessment of actual writing ability; however, it may be questioned why the rating has been limited to these three factors. The task may very well lack construct validity and as we shall discuss later, the accurate assessment of writing ability may require rather more factors than are demanded here. Furthermore, one disadvantage of analytic scoring, especially when used in the absence of holistic scoring, is that it may, in the words of Hughes, "divert attention from the overall effect of the piece of writing" (Hughes, 2007, p. 103). Indeed, he goes on to say that "overemphasis on such mechanical features as spelling and punctuation can invalidate the scoring of written work" (Hughes, 2007, p. 33). Finally, the teachers are asked to provide a score out of 25 for each examinee based on these five criteria. The designer of this rating system may have simply assumed that the teachers would simply mark each criterion out of 5, but this may not be the case. Some teachers may put more weighting on the first three criteria at the expense of the final two. This will seriously jeopardize the reliability of the test scoring by creating huge variance in the test-takers' scores. As stated by Bachman and Palmer, "if some raters rate more severely than others…the scores obtained could not be consid-

ered to be reliable" (Bachman & Palmer, 1996, p. 20).

In contrast to the first approach, the second marking scheme seems far better in terms of reliability and validity. First of all, in order to ameliorate the negative effects of subjectivity, the scripts are marked by two teachers of English. As Hughes states with regard to reliability, "as a general rule, and certainly where testing is subjective, all scripts should be scored by at least two independent scorers" (Hughes, 2007, p. 50). Another important factor here is that the scoring is carried out independently and marks are only compared after the process has been completed. The fact too that they are both English teachers would seem to ensure a certain level of competence and experience. One possible criticism is that there is little indication as to the level of training that these two scorers have received. In the words of Hughes, ' [training] is especially important where scoring is most subjective" (Hughes, 2007, p. 49) and further suggests that scores should be analyzed for deviancy from the norm. In this case, with only two scorers, the need for training would be even more essential. As to the method of scoring, this approach includes both holistic and analytic scoring. This is far better than the solely analytic approach employed in the first marking scheme. First of all, teachers quickly read the script in order to gain an overall impression. This is important because a piece of writing is so often more than merely the sum of its parts. It is a valid tool in assessment in that it is actually assessing the students' overall ability to write without necessarily breaking their scripts down into their constituent parts. Following this, the teachers then mark the scripts based on a number of clearly stipulated criteria. This marking scheme has previously been 'developed to link with identified levels of proficiency' and thus we can probably assume it has been tried and tested. The criteria include three of those adopted in the

first marking approach (spelling, punctuation and grammar) as well as vocabulary, cohesion and sentence structure. This ought to improve reliability because, as Hughes states, "the more scores for each candidate, the more reliable should be the final score" (Hughes, 2007, p. 94). It would seem rational that the more criteria a student is marked on the greater the reliability of the marking. However, one does have to strike a balance between reliability and practicality; as Harmer suggests, a scoring key that includes "a profusion of criteria may make the marking of a test extremely lengthy and cumbersome" (Harmer, 2007, p. 309). Granted, analytic scoring is time consuming, but the advantage of using both multiple analytic scoring as well as holistic scoring is that higher accuracy can be obtained. In addition to this, as Hughes states, "significant discrepancies" between the two totals can thus be investigated should they arise and this should guard against the dangers of concentrating too much on the different aspects which "may divert attention from the overall effect of the piece of writing" (Hughes, 2007, p. 103). This can only add to the validity of the task. The scorers were adequately guided in their scoring by the use of descriptive criteria. The detailed scoring key, certainly for that of cohesion, seems to guide the markers effectively by stating unambiguously what is deemed as acceptable. This is a big improvement on the first marking approach which clearly fails to give clear guidelines as to how to consistently score the language features. Finally, this second marking approach is superior to the first in that the final scores are then compared and any inconsistencies are mutually analyzed by referring back to the marking criteria. This would seem to be a better solution to that suggested by Hughes who believes that such matters should be settled by a "third, senior colleague who compares the two sets of scores and investigates discrepancies" (Hughes, 2007, p. 50). It is far better, in my opinion, for the two scorers to discuss their differences together, as it

could help to iron out any possible ambiguities in the marking scheme.

This paper has endeavored to look at the theory and practice behind the design and application of a good language test. In particular it has highlighted the crucial areas of reliability, validity and authenticity as the foundations of such a test. It also considered the concept of washback as an agent for encouraging good teaching practice and boosting the confidence of students.

References

Bachman, L., & Palmer, A. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: Oxford University Press.

Brown, H.D. (2004). *Language Assessment: Principles and Classroom Practices*. New York: Pearson Education.

---. (2007). *Teaching by Principles: An Interactive Approach to Language Pedagogy*. 3rd ed. New York, NY: Pearson Education, Inc.

Harmer. J. (2007). *The Practice of English Language Teaching*. 4th ed. Essex: Pearson Education Limited.

Henning, G. (1987). *A Guide to Language Testing: Development, Evaluation, Research*. Boston, Mass: Heinle & Heinle.

Hughes, A. (2007). *Testing for Language Teachers*. 2nd ed. Cambridge: Cambridge University Press.

Mangubhai, F. (2008). *Language Testing*. University of Southern Queensland.

O'Neill, S. & Gish, A. (2008). *Teaching English as a Second Language*. Victoria: Oxford University Press.