CAA VALIDATION IN NIECELT AND SOME PEDAGOGICAL ISSUES IN QUESTION DESIGN AND CONTENT OF AN INNOVATIVE CLOZE

Anthony Seow, Chew Lee Chin & Luo Guanzhong

CAA Validation In NIECELT And Some Pedagogical Issues In Question Design And Content Of An Innovative Cloze

Anthony SEOW, CHEW Lee Chin & LUO Guanzhong
National Institute of Education
Nanyang Technological University
1, Nanyang Walk
Singapore 637616

Abstract

The National Institute of Education (NIE) in Singapore has developed a comprehensive testing software known as "NIE Computerised English Language Test" or, simply, NIECELT, that is capable of administering a test or a number of tests to any specified number of examinees at the same time or at different times. Inherent in NIECELT are a number of interactive test questions which allow the examinees to craft some of their answers in response to a set of questions that assess the examinees' proficiency in language structure, grammar, vocabulary, reading-comprehension and the use of language in context.

An issue challenging NIECELT as a computer-assisted assessment tool pertains to the defensibility of inferences made from the obtained test scores. Related empirical questions include: 1) How well does performance on NIECELT reflect examinees' language proficiency compared to that measured by the English Language GCE 'O' level examination? 2) What are the relationships among the different test questions in NIECELT presented to examinees?

This paper attempts to provide evidences of both the internal and external structures of NIECELT. Implications of the findings are discussed in terms of the pedagogical issues in question design and content of one interactive question type – a modified cloze procedure that tests essentially language in context. The innovation in question type in the cloze will also be clarified.

Keywords

Test Validation, Cloze

Introduction

With increased availability of computers and a broader access to computers, computer-assisted assessment (CAA) has become a reality. Compared to the conventional paper-administered testing, CAA uses the computer as a medium of test administration and this opens up opportunities for novel question types, design and content. But as with all testing, CAA or otherwise, test validity is the most important consideration in test evaluation (AERA, APA & NCME, 1985). Popham (1995) asserts that the validity of a test is measured by the defensibility of score-based inferences. In other words, test validation is a process of accumulating evidence to support a particular inference for test scores. Messick (1989), and Linn, Baker, and Dunbar (1991) have suggested several types of validity evidence, and each type of evidence provides some usefulness for validating the interpretation and use of test results. Shepard (1993) suggests that validity evidence be prioritized for an assessment practice, that is, what is intended for interpreting and using the test results in a particular situation.

Background

English language proficiency tests are routinely administered to potential trainee teachers who seek admission and placement to various programmes at the National Institute of Education (NIE). At present, four major tests are used and these tests aim at assessing various language skills, including speaking, listening, reading, discourse and grammar and vocabulary competence. But the paper and pencil mode of testing currently in use is complex and consumes considerable staff time and energy, largely because of the existence of tests that overlap in several areas, the frequency of testing, and the large numbers of candidates that take some of the tests. An overall research effort was initiated in 1997 to streamline NIE language testing by incorporating the use computer technology. The research project team has developed a testing software prototype known as “NIE Computerised English Language Test” or, simply, NIECELT, that is capable of administering a test or a number of tests to any specified number of examinees at the same time or at different times. Inherent in NIECELT are a number of interactive test questions which allow the examinees to craft some of their answers in response to a set of questions that assess the examinees’ proficiency in language structure, grammar, vocabulary, reading-comprehension and the use of language in context.

Purpose

The present study was to determine the empirical validity of NIECELT as a computer-assisted assessment for English language. If inferences made from the obtained test scores are defensible, then there is better understanding and confidence in using this new method for assessing trainee teachers’ English language proficiency at NIE.

The cloze procedure (or, cloze) within NIECELT is highlighted if only because it serves this paper well to illustrate, through one but nevertheless important test component of NIECELT, the pedagogical issues and innovativeness of its question

design. A cloze is “a procedure in which deletions are made in a text, usually of single words selected pseudo-randomly (e.g. regular deletions of every seventh word, or some other number), and test takers are asked to supply the missing words. Other associated variations include rational deletion procedure and C-tests” (Allison, 1999:230). The innovative cloze in NIECELT is a modified cloze with rational deletions that follows certain pedagogical and testing principles. The NIECELT cloze will be discussed later in this paper.

Method

Participants

The 84 participants in a recent pilot study of NIECELT were trainee teachers enrolled in three different academic programmes in NIE, namely, the diploma, the degree and the postgraduate. All of the participants were volunteers and they had no prior experience with computer-assisted assessment (CAA).

Instruments

The NIECELT: The CAA instrument in the original NIECELT is a 100-item test aimed at assessing the trainee teachers’ language proficiency in four areas of language skills: 1) Grammar 2) Vocabulary 3) Reading-Comprehension 4) Cloze.

Background Information Questionnaire: This is a survey questionnaire designed to collect some background information of the participants. This includes gender, age group, educational experience and highest academic qualifications, English language grades obtained at the GCE ‘O’ and ‘A’ levels, and languages spoken at home.

Procedure

The trainee teachers were contacted for a CAA two weeks prior to the pilot study, which was conducted in August 2000. The participants had two hours to complete the test. A survey questionnaire was also administered just before the test to collect some background information of the participants.
Descriptive statistics were used to yield the participants’ test scores obtained on NIECELT and their English Language (EL) grades obtained in their GCE ‘O’ level examination.

The relationships among the CAA components in NIECELT were determined using the Pearson’s product-moment correlation. This statistical technique was also used to elicit evidence of criterion-related relationship for the NIECELT, which was established by comparing the participants’ test results in NIECELT with the participants’ English Language results in the GCE ‘O’ level examination.

Results

The means and standard deviations for the number of correct responses to the test items in the various sub-tests in the CAA NIECELT are presented in Table 1.

On the whole, this CAA NIECELT seems a "difficult" test if the mean scores alone as seen in Table 1 are used as a convenient yardstick. Particularly revealing are *Paraphrase* (Mean=3.4 or 34.0%) and *Errors and Corrections II* (Mean=0.6 or 8.6%) subsumed under Grammar (Mean=12.6 or 36.0%), and *Opposites* (Mean=1.8 or 36.0%) categorized under Vocabulary. However, the Vocabulary domain (Mean=18.1) on the whole appreciated to 64.6% on account of the much stronger mean scores bolstered by *Word Choice* (75.0%) and *Filling in the Blanks* (68.7%).

Table 1.
Means, standard deviations (SD) and percentage mean scores of correct responses observed in the sub-tests in NIECELT

Sub-tests in NIECELT	No. of test-items	Mean scores (SD)	Mean percentage scores
1. NIECELT	100	50.0 (10.3)	50.0%
2. Grammar	35	12.6 (4.1)	36.0%
a. Word Unscrambling	10	4.8 (1.7)	48.0%
b. Paraphrase	10	3.4 (2.3)	34.0%
c. Errors & Correction I	8	3.7 (1.9)	46.3%
d. Errors & Correction II	7	0.6 (0.8)	8.6%
3. Vocabulary	28	18.1 (3.5)	64.6%
a. Word Choice	8	6.0 (1.4)	75.0%
b. Opposites	5	1.8 (1.1)	36.0%
c. Filling in the Blanks	15	10.3 (2.6)	68.7%
4. Reading-Comprehension	12	6.3 (2.4)	52.5%
5. Cloze	25	12.9 (4.3)	51.6%

The apparent difficulty of NIECELT could be ascribed partly to the test content and partly to the fact that the test-takers were probably unfamiliar with the CAA mode of testing. The length of the test, i.e. 100 questions spread over a number of different test formats for which the test-takers would require extra time to get accustomed to, might itself be a debilitating factor.
If we began to speculate about the likely sources of content difficulty in regard to Paraphrase, Errors and Corrections II and Opposites, we could come up with some plausible interpretations.

In the Paraphrase sub-test, with a given structure, the test-takers were required to re-construct a new structure, using a beginning cued word and keeping to the sense

of the given structure. In essence, this means that we are testing the test-takers' ability to perceive "systematic correspondences" between one structure and another (see e.g. Quirk et. al, 1989:57). To do this successfully, the test-takers would need to demonstrate their ability to understand "the relation between grammatical choice and meaning" (ibid) in converting from one structure to another that is closely parallel in meaning. This seemed to be not an easy task for many of the test-takers.

The Errors and Corrections II sub-test assesses the ability of the test-takers to not only know how to correct errors but also to be able to first identify those phrases (or clusters of words) that contain errors. These errors are themselves very subtle in that they are those errors commonly produced by L2 learners (e.g. to request for*) and are thus not so easily recognizable as erroneous. This sub-test in NIECELT was observed to be the most difficult to handle. This explains why its mean score was a meagre 0.6 (or 8.6%).

The Opposites sub-test, although on a familiar four-option multiple-choice question format was, surprisingly, difficult. One explanation could be that the lexical items tested were rather uncommon words and that the four highly plausible options to each question were challenging.

Imposing an interpretation as to why these three sub-tests had low mean scores is not necessarily an apology for having created poor sub-tests. Rather, it had alerted us to the fact that these tests might in fact be excellent test items for discriminating the really good test-takers from the weaker ones. At the moment, it remains mere speculation, and more research needs to be done on this. NIECELT is undergoing periodic revisions to make it work even better after taking into account, question design, content and the test-takers' feedback on the degree of operational ease of the individual sub-tests.

Table 2, which shows the Pearson product-moment correlations between the EL test scores of the test-takers in NIECELT and in their GCE 'O' level examination, demonstrates that participants who performed well in the national GCE examination are also more likely to perform well in the CAA NIECELT ($r = .47$, significant at the .01 level). This is particularly true of the GCE EL grades vis-a-vis the Grammar and Vocabulary scores respectively in NIECELT, where the correlations are moderately high ($r = .36$ to $.43$). There is also a direct relationship between the EL exam grades and the Reading-Comprehension scores ($r = .28$) and the Cloze scores ($r = .27$). Of some concern is the dismally weak association between the EL exam grades and the respective scores for Paraphrase, Errors & Correction II and Opposites in NIECELT. These three sub-tests were earlier observed to have the lowest mean percentage scores of correct responses.

Table 2.
Product-moment correlations between EL performance in NIECELT and EL grades obtained in the GCE 'O' level examination

Sub-tests in NIECELT	Correlation Coefficients derived from NIECELT scores and GCE "O" level EL grades
1. NIECELT	.47 **
2. Grammar	.43 **
a. Word Unscrambling	.42 **
b. Paraphrase	.11
c. Errors & Correction I	.37 **
d. Errors & Correction II	.14
3. Vocabulary	.36 **
a. Word Choice	.33 **
b. Opposites	-.03
c. Filling in the Blanks	.33 **
4. Reading-Comprehension	.28 **
5. Cloze	.27 *

** Correlation is significant at the 0.01 level (2-tailed)
* Correlation is significant at the 0.05 level (2-tailed)

Table 3.
Intercorrelations among the sub-tests in NIECELT

Sub-tests in NIECELT	1	2	2a	2b	2c	2d	3	3a	3b	3c	4	5
1. NIECELT	1.00	.78**	.67**	.47**	.41**	.29**	.70**	.56**	.25*	.56**	.60**	.74**
2. Grammar		1.00	.63**	.67**	.68**	.32**	.42**	.37**	.12	.32**	.29**	.41**
a. Word Unscrambling			1.00	.10	.25*	.24*	.51**	.45**	.19	.38**	.37**	.38**
b. Paraphrase				1.00	.19	-.05	.24*	.14	.10	.22*	.15	.21
c. Errors & Correction I					1.00	.13	.08	.12	.01	.04	.04	.25*
d. Errors & Correction II						1.00	.19	.26*	-.07	.15	.19	.13
3. Vocabulary							1.00	.68**	.37**	.68**	.35**	.26**
a. Word Choice								1.00	.16	.34**	.24**	.29**
b. Opposites									1.00	.01	.11	.13
c. Filling in the Blanks										1.00	.31**	.15
4. Reading Comprehension											1.00	.32**
5. Cloze												1.00

** Correlation is significant at the 0.01 level (2-tailed)
* Correlation is significant at the 0.05 level (2-tailed)

Table 3, which shows the inter-correlations amongst the subtests in NIECELT, reveals the internal structure of NIECELT. Moderately high positive correlations, ranging from .60 to .79, were obtained between the respective scores of the four skills tested – i.e. Grammar, Vocabulary, Reading-comprehension and Cloze - and the overall NIECELT scores. The relationships thus manifested suggest that each skill tested contributes positively towards the credibility of each of these language skills tested in NIECELT.

An examination of the four Grammar subtests (2a to 2d) reveals strong relationships for Word Unscrambling, Paraphrase, and Errors & Corrections I, with the overall Grammar score, with coefficients ranging from .63 to .68. But the sub-test of Errors & Corrections II shows a weaker relationship with Grammar (r = .32). Recognized, nevertheless, as a good sub-test by the NIECELT project team, Errors and Corrections II could, however, be improved further to follow more closely the tradition of clause analysis found in Quirk, et. al (1989), for example.

The three Vocabulary sub-tests (3a to 3c) reveal strong relationships between the overall Vocabulary score and Word Choice (r =.68) and Filling in the Blanks (r =.68) respectively, but the relationship between the Vocabulary score and Opposites is weaker (r =.37).

A closer examination of the relationships among the tests in NIECELT reveals some interesting information. A moderately positive correlation obtains between the Grammar and the Vocabulary test scores (r = .42). For both the Cloze and Reading-comprehension tests, low to moderate positive correlations, ranging from .26 to .41, were obtained with Grammar, Vocabulary, and with each other. This observation provided us the motivation to improve on the Cloze in NIECELT in a later revision following the pilot study, to make it a more effective test of language in context that incorporates the testing of the skills of grammar, vocabulary and reading-comprehension. The challenge for us then was to decide what to include in the revised Cloze that is now more innovative from the CAA perspective and more credible from the viewpoint of testing principles.

Discussion

The results reveal some interesting evidences of both the internal and external structures of NIECELT.

Finding 1:

Assessment results from CAA NIECELT show satisfactory consistency with the results from GCE ‘O’ level EL examination. An earlier validation study of a paper-administered EL proficiency examination conducted with trainee teachers at NIE (Chew et al., 1997) found a similar relationship with the GCE ‘O’ level examination. NIECELT, however, assesses a wider range of language skills, made possible because of the more objective CAA mode of testing.

Finding 2:
Overall, the four domains of Grammar, Vocabulary, Reading Comprehension and Cloze tested in NIECELT contribute positively towards assessing the EL proficiency skills of trainee teachers. However, it is noted that the weaker relationships of some skills in the sub-tests could be due to inherent problems of question design and content.

According to Messick's (1989) conception of validity, it is also important to elicit evidences of the consequential basis of test interpretation and test use.
For NIECELT, this would mean understanding and checking on any unintended consequences of the computer-assisted assessment and resolving them. However, for the purpose of this paper, only the cloze is highlighted for special mention.

Question Design And Content Of An Innovative Cloze

In question design, the Cloze in NIECELT has taken care to adhere closely to the general principles of L2 communicative language testing. In particular, it makes "use of authentic texts" and assesses "the (learner's) ability to integrate grammatical, lexical contextual, and pragmatic knowledge in test performance" (McNamara, 2000:16-17). If indeed, as evidenced earlier, there was a direct positive link between the Cloze and Grammar, Vocabulary and Reading-comprehension respectively, then it makes pedagogical sense to ensure that the rational deletion of words in the Cloze in NIECELT should consciously aim to cultivate the development and assessment of grammatical, vocabulary and reading skills in the test-takers. In other words, the test designer should make informed decisions about which words in the target cloze passage to delete precisely. For this reason, rather than relying on the traditional "true" cloze where the deletion of words in the passage is made (in a sense, quite mechanically) at regular intervals, a modified cloze is preferred.

There is an abundance of research evidence to date to suggest that the cloze is "an invaluable means of assessing a student's all-round command of the English Language, in the *grammar and usage* aspects, in *vocabulary*, in general knowledge and in experience" (Oei, 1988. See also Garman & Hughes, 1983), as well as "a meaningful way of helping *reading* in the classroom" (Rye, 1982; Weir, 1995). In the latter case, Garman & Hughes (1983:Introduction, vii) claim even further that the cloze procedure as a teaching tool is "much more economical than the traditional reading passage with associated questions."

The big question is, of course, what should constitute the content for the cloze. The important content trademark of the innovative Cloze in NIECELT is that the authentic text used attempts to test a wholesome bundle of language skills that relate to grammar, vocabulary, reading-comprehension, word collation and grammatical and lexical cohesion. Thus, specific types of words relating to these language skills being tested are targeted for rational deletions in the Cloze in NIECELT. An example of such a rational cloze is seen in Figure 1.

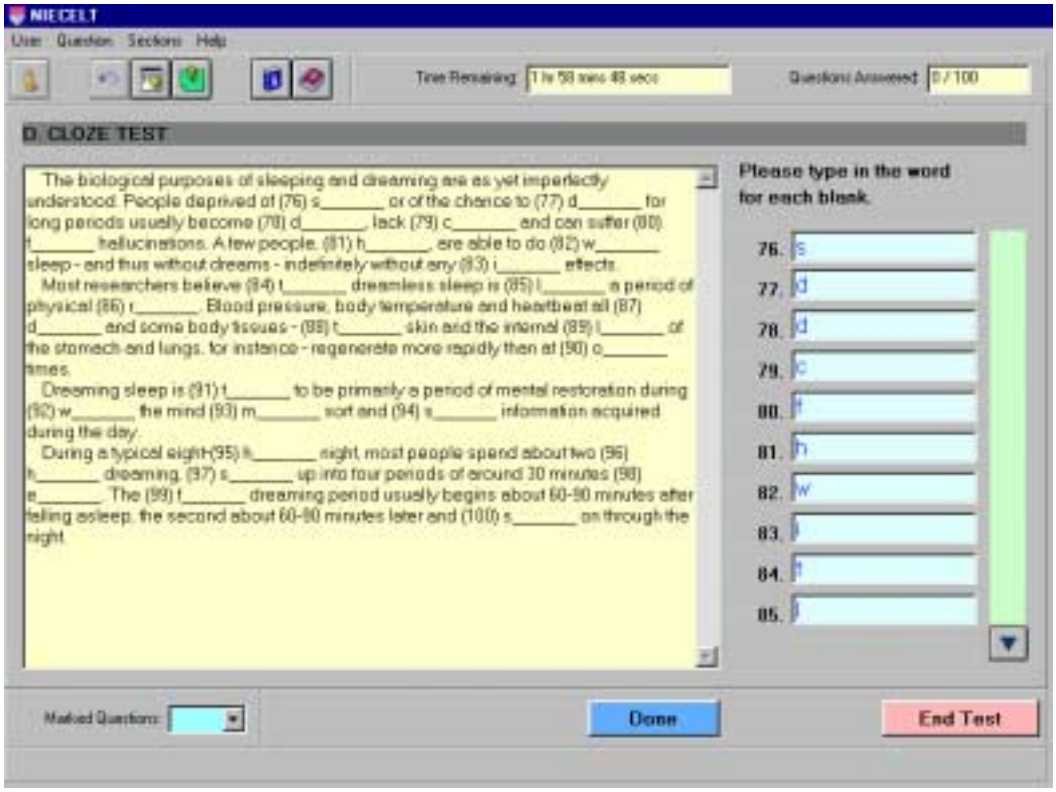

Figure 1. A Typical Rational Cloze In NIECELT

Central to the NIECELT Cloze is text cohesion. Grammatical and lexical cohesion assessed in the Cloze includes the categories mentioned in Halliday & Hasan (1976): *reference* (e.g. pronominals, demonstratives, definite article, comparatives), *substitutions*, *ellipses* (e.g. nominal, verbal, clausal), *conjunctions* (e.g. enumeration, exemplification. comparison/contrast, chronology, cause/effect) and *lexical cohesion* (e.g. repetition, superordination).

The **innovativeness** in the Cloze in NIECELT is observed in several respects. For one, we have, in an authentic text, words deleted in well-defined locations, which account for testing specific language skills in context. The first letter of every deleted word is left intact. This is a variation of the C-test (Weir, 1995:80), and its efficacy lies in the fact that the target answer to each blank is the *original* word, that is, the word as used by the author in the original passage. It seems to us that there is much more pedagogical and testing value in calling for the original word rather than any other suitable alternative simply because the original sense intended by the author is wholly retained.

What is perhaps more remarkable is that immediately after the test-taker has supplied an answer to a blank (by typing in the word in answer column – See Figure 1), the intended answer actually appears in the blank in the passage! The text meaning in the cloze passage is thus progressively built up as the test-taker completes the test from one blank to another. This innovative feature of the NIECELT Cloze is highly desirable in a testing as well as in any learning situation.

Using the NIECELT Wizard in the construction of a Cloze, the test designer - and we have in mind the ever-busy classroom teacher as well – has the benefit of using any previously prepared text (e.g. on *Microsoft Word*) to be cut-and-pasted electronically onto the test template for immediate use in the construction of the Cloze. What the test designer (or the teacher) needs to do with the prepared text is simply to highlight any word they want deleted and with the click of the “Set As Blank” button on the NIECELT Cloze template, the desired blank prefixed with a question number is sequentially created. The answer to this blank, automatically programmed into the computer, then appears in the answer list in the computer. This innovative idea of NIECELT is illustrated in Figure 2 below.

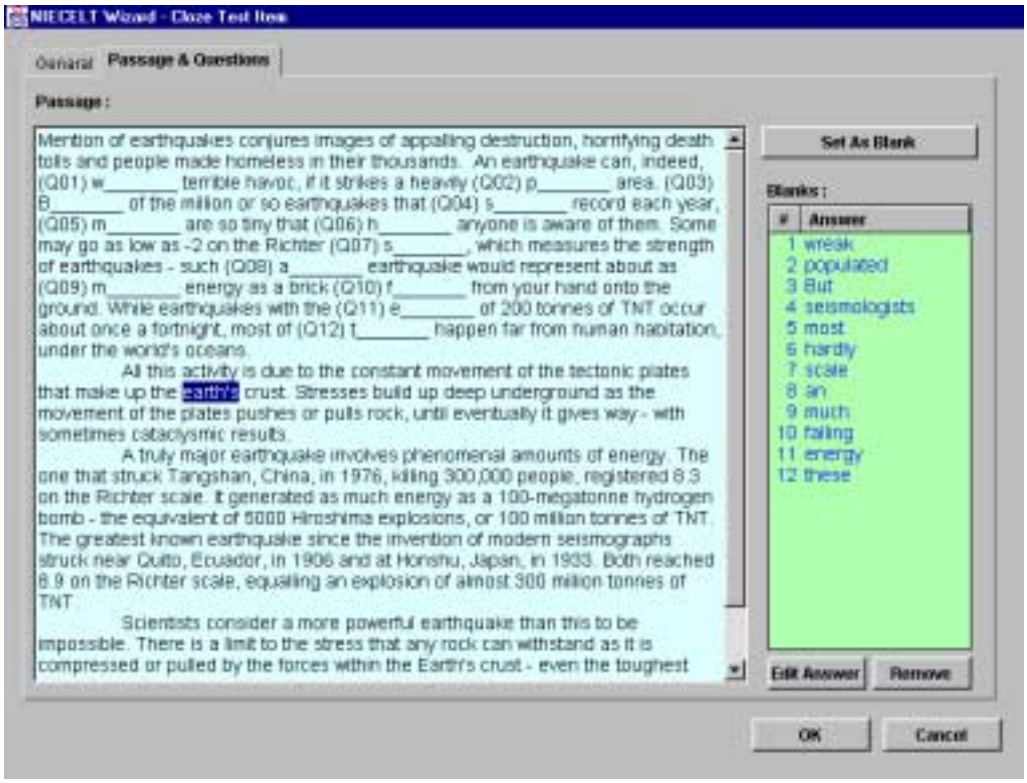

Figure 2. Creating A New Cloze

Should the test designer decide to put a deleted word back into the passage, all is not lost since at the activation of the “Remove” button, the original word is returned to the passage and all signs of its having been removed before and a numbered blank created in its place are obliterated.
The NIECELT Cloze – indeed the whole NIECELT system with its Wizard – is a cinch for those who wish to use it for CAA.

References

Allison, D. (1999). *Language testing & evaluation*. Singapore: Singapore University Press.

Chew L. C., Hsui, V. Y., & Seow, A. (1997). *An assessment of some test-criterion relationships of English language tests*. A paper presented at the Annual Conference of the Educational Research Association, November 1997, Singapore.

Garman, M. & Hughes, A. (1983). *English Cloze Exercises*. Oxford: Basil Blackwell.

Halliday, M.A.K. & Hasan, R. (1976). *Cohesion in English*. London: Longman

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. Educational Researcher, 20(8), 5-21.

McNamara, T. (2000). *Language Testing*. Oxford: Oxford University Press.

Messick, S. (1989). Validity. IN R. L. Linn (Ed.), Educational measurement (3rd ed.) (pp. 13-103). New York: Macmillan.

Oei, S.K. (1988). Singapore: New ERA Publishers.

Popham, W.J. (1995). *Classroom assessment*: What teachers need to know. Boston: Allyn & Bacon.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1989). *A comprehensive grammar of the English Language*. London: Longman

Rye, J. (1982). *Cloze procedure and the teaching of reading*. London: Heinemann.

Shepard, L. E. (1993). *Evaluating test validity*. Review of Research in Education, 19, 405-450.

Weir, C. (1995). *Understanding & developing language tests*. New York: Phoenix ELT.