دولة الإمارات العربية المتحدة
كلية الدراسات الإسلامية والعربية بدبي

مجلة كلية الدراسات الإسلامية والعربية

مجلة علمية محكمة

اقرأ في هذا العدد

العدد السابع والثلاثون

البريد الإلكتروني iascm@emirates.net.ae

الموقع الإلكتروني www.islamic-college.co.ae

1430هـ /2009 م

اثنان وعشرون عاماً من العطاء

مَجَلَّــة
كُلِّيَّةِ الدِّراساتِ الإسلاميَّةِ والعربيَّةِ

مجلة علمية محكَّمة

نصف سنويَّة

العدد السابع والثلاثون

جمادى الآخر ١٤٣٠ هـ - يونيو ٢٠٠٩م

ردمد: ١٦٠٧-٢٠٩X

تفهرس المجلة في دليل أولريخ الدولي للدوريات تحت رقم ١٥٧٠١٦

المحتويـــــــات

ملخص البحث

دراسة تحليلية لفاعلية الاختبار

د. خالد الخاجه

د. مريم بيشك

من أهداف أي إختبار هو قياس فاعلية التعليم أو التعلم أوكلاهما مقارنة بالأهداف المعلنة للنظام التربوي أو الإرشادي. ولكون اللغة نظام معقد وغير ملموس فليس من السهولة بإمكان قياسها قياسا مباشرا وواضحا. فقد فشلت عموما نظريات اختبار التواصل اللغوي التحليلية والتكاملية السيسولغوية في طرح طرائق اختبار ناجحة معترف بها عالميا.

لذا يهدف هذا البحث إلى التعرف على بعض المشاكل النظرية والعملية (للتيسول) والتي تقدم توصيات لواضعي الاختبارات التي من شأنها أن تراعي كل من المتعلم والمادة التعليمية.

Madison, H. 5. (1983). *Techniques in Testing*. Oxford: Oxford University Press.

Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing, Vol. 26, pp. 75 - 100.*

Moller, A. (1975). Validity in Proficiency Testing, *ELT Documents* 3, British Council. Morrow, K.E. (1979). 'Communicative Language Testing: Revolution or Evolution'. In C.

J. Brumfit, and K. J. Johnson (eds.) *The Communicative Approach to Language Teaching*. Oxford: Oxford University Press.

Mullen, K. (1979). More on doze tests as tests of proficiency in English as a second language. In Briere, E. & F. Hinofotis (eds.), *Concepts in language testing*: Washington, D.C.: Teachers of English to Speakers of Other Languages.

Oller, J. (1973). 'Cloze Tests of Second Language Proficiency and What They Measure'. *Language Learning*: 23: 105-18.

----------. (1979). *Language tests at school*. London: Longman.

---------- (1987). Practical Ideas for Language Teachers from a Quarter Century of Language Testing. *English Teaching Forum*: 42-46.

Read, J.A.S. (1981). (ed.) Direction in Language Testing, *RELC Anthology*, Series 9. SEAMEO Regional Language Centre, Singapore.

Rivera, C. (1984). *Communicative Competence Approaches to Language Proficiency Assessment: Research and Application*. Clevedon: Multilingual Matters.

Romiszowski, A. (1986). *Developing Auto-Instructional Materials*. London: Kegan Page.

Saleemi, A. P. (1988). Language Testing: Some Fundamental Aspects. *English Teaching Forum*: 2-6.

Savignon, 5. (1972). Teaching for Communicative Competence: A Research Report. *Audio-Visual Language Journal*, 10 (3): 153-162.

Spolsky, B. (1985). What Does It Mean to Know How to Use a Language? An Essay on the theoretical basis of language testing. *Language Testing*, 2 (2): 189-199.

Stub, J. B. and G. Tucker (1974). 'The Cloze Test as a Measure of English Proficiency'. *Modern Language Journals*, 58 (5/6): 239-4 1.

Taylor, W. (1953). 'Cloze Procedure: A New Tool for Measuring Readability'. *Journalism Quarterly*, 30: 425-436.

Underhill, N. (1982). The Great Reliability - Validity Trade-off. Problems in assessing the productive skills. In B. Heaton (Ed.), *Language Testing*. London: Modern English Publications Limited.

Weir, C. J. (1988). *Communicative Language Testing*. London: University of Exeter.

-------- (1993). *Understanding and Developing Language Tests*, London: Prentice Hall International.

References

Alderson, J. (1978a). The Effect of Certain Methodological Variables on Cloze Test Performance and its Implication for the Use of the Cloze Procedure in E.F.L. Testing. *In 5th International Congress of Applied Linguistics, Montreal*. Montreal.

------------ (1 978b). *The Use of Cloze Procedure with Native and Non-native Speakers of English* [Ph.D. dissertation]. Edinburgh: University of Edinburgh.

Bartz, W. (1979). 'Testing Oral Communication in the Foreign Language Classroom'. In *Education Theory and Practice*, 2/17.Arlington, Va: Center for Applied Linguistics.

Canale, M., & Swain, M. (1980). 'Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing'. *Applied Linguistics*, 1(47): 1-47.

Carroll, B. J. (1961) 'Fundamental Consideration in Testing for English Language Proficiency of Foreign Students'. In Allen, H.B. and R.A. Campbell (Eds).

Chaplin, E. (1970). *The Identification of Non-Native Speakers of English Likely to Under-*

Achieve in University Courses Through Inadequate Command of the Language [Ph.D. Thesis]. Manchester: University of Manchester.

Cohen, A. D. (1980). *Testing Language Ability in the Classroom*. Rowley, Mass.: Newbury House Publishers.

Cooper, C. (1977). Holistic Evaluation of Writing. C. Cooper & L. Odell (eds.), (pp. 3-31). Urbana, IL: National Council of Teachers of English.

Darnell, D. (1968). *The Development of an English Language Proficiency Test of Foreign Students Using a Clozentropy Procedure*. Boulder, CO: University of Colorado.

Davies, A. (1978). Language Testing. In Kinsella, V. (ed.) *Language Teaching and*

Linguistics Abstracts, Vol. 2: No. 3 & 4: 127-159. Cambridge: Cambridge University Press.;0]

Ediger, M. (1988). 'Evaluating Learner Progress in Reading'. *English Language Journal*, 19, 1988: 1-4.

Farhadi, H. (1979). 'The Disjunctive Fallacy Between Discrete Point and Integrative Tests'. *TESOL Quarterly*, 13 (3): 350.

Fulcher, G. and Davidson, F. (2009). Test architecture, test retrofit. *Language Testing, vol.*

26, pp. 123 - 144.

Green, J. A. (1975). *Teacher-Made Tests*. New York: Harper & Row, Publishers.

Hale, G. A., C. W. Stansfield & R. P. Duran (1984). 'Summaries of Studies Involving the Test of English as a Foreign Language 1963-1982. In *Research Report. Educational Testing Service*, Princeton: 17-149.

Haskell, J.F. (1976). Using Cloze to Select Reading Material. *TESOL Newsletter*, 10(11):

15-16.

Heaton, J. (1975). *Writing English Language Tests*. London.: Longman.

Hinofotis, F. A. (1976). *An Investigation of the Concurrent Validity of Cloze Testing as a Measure of Overall Proficiency in English as a Second Language* [Ph.D. Dissertation]. Southern Illinois University.

Hirvala, A. (1989). Using Cloze Passages for Instructional Purposes. *TESL*

Reporter.

Ilyin, D. (1976). *Ilyin Oral Interview*. Rowley, Mass.: Newbury House.

Jacobs, H., S. Zinkgraf, D. Wormuth, W. F. Hartfield, & J. Hughey, (1981).*English Composition Program. Testing ESL Composition: A Practical Approach.* Rowley, Mass.: Newbury House.

Johnson, R. K. (1982). Questioning Some Assumptions about Cloze Testing. In B. Heaton (Ed.), *Language Testing*. London: Modern English Publications Limited: 59- 63.

Jonz, J. (1990). 'Another Turn in the Conversation: What Does Cloze Measure'? *TESOL*

Quarterly, Vol. 61, pp. 3-23.

comprehension is required. Although some teachers use doze tests to assess reading comprehension, one must remember that such tests measure overall proficiency. Tests, such as multiple-choice, though relatively easy to design, administer, and score do not demand active and original production on the part of the testee; they expose the students to errors which are imprinted in the brain; they may also falsify the interpretation of the score inasmuch as the scorer does not know how much thinking and how much guessing has gone into answering the test.

Finally, it has often been observed that overtesting is likely to produce diminishing returns. Testing should not be an end in itself It is, first, a teaching instrument; second, an assessing tool. Quizzes motivate without paralyzing because they are not expected to have much weight and because it has been explained to the student that their purposes is not to grade, to judge, but to reinforce learning and guide the teacher as to what should be taught and how it should be taught.

Testing is a skill and can, therefore, be learned. It is also an art, which demands not only technical competence, but psychological, social, and cultural sensitivity as well.

learners. There is no one best method for all or most people: “Language users really do invent and internalize rule systems that, as Chomsky and others have consistently maintained, are generative in character” (Oller 1987: 48). The rates and the manners of internalization and invention are individual characteristics.

Underhill (1982) remarked that “the multiple-choice item is a thoroughly unrealistic measure of language performance. It does not reflect actual language use.” Though this is doubtlessly true, it is equally true that is does measure specific linguistic competencies, and if this happens to be one of the tester’s objectives, multiple-choice may be a good choice.

It does not assess face/content validity, which can be measured only through qualitative criteria, as an analytic method, it does possess much greater reliability than synthetic tests—a consideration in a test “package”.

Underhill (1982), having analyzed the various methods for assessing the productive skills of foreign-language learners, concludes that “what kind of test we used should be determined pragmatically by the purpose for which test you use, the resource you have available for construction, administration and marking, and what you intuitively feel will have the highest face/content validity for testees and testers alike.”

Cohen (1980: 1) notes that “One reason for testing is to promote meaningful involvement of students with material that is central to the teaching objectives of a given course.” Tests provide feedback is an imperative of learning in fact; there is no learning without feedback—by definition.

To test for reading comprehension teachers often report to true-false questions or to the matching technique. They do so because they believe this is an easy way out of the intricate design of tests: a True-False item seems easy to prepare and, at least for the beginning student, they demand less intellectual effort than the multiple-choice type of technique. Madison (1983) points out, though, that “one problem with the true-false question is that the student might simply guess the right answer. If concerned about this, you can make a correction for guessing: Just subtract the number wrong from the number right. This is their new score.” As a matter of fact, Madsen adds that a guessing correction can also be made for multiple-choice tests: “The correction is made by dividing the number of items wrong by the number of distracters and subtracting this from the number of correct answers.”

As to the matching technique, it too is simple to construct and to respond to: students match material in the passage with material in the question. Obviously, though providing practice in responding to questions, little if any and that any "use of language to represent meaning is a potential language test". He thus advocated meaning-oriented language tests as opposed to surface-oriented language tests. These letter tests are expression of the 1940-50s structural linguistics' discrete-point testing which emphasized structural patterns, sounds, and word forms. Meaning-oriented testing, on the other hand, focus on the performance of communicative tasks. Indeed, authentic language use always involves a linking of elements of text (speech included) which the ongoing stream of experience. This process has been called pragmatic mapping.

Pragmatic mapping may be defined simply as an intelligent and articulate connecting of facts with text, or of experience with language. In fact, if the dictated text is uttered at a conversational rate a burst of three to seven words (or more) so as to present a challenge to short-term memory, dictation turns out to be a highly effective way of testing a person's ability to follow spoken version of a given text with comprehension (Oller 1987).

Actually, dictation is "one of the easiest tests to use, and it gives very good information on the student's language ability. But this is true only if you prepare it right, and score it right" (Madison 1983: 112). Madison (Ibid: 114) suggests that the best way to score a dictation test is to deduct one point for each error. We recommend this even if you are counting off for spelling and punctuation errors. It might seem fairer to take several points off for serious errors and fewer points off for less serious errors. But much practical experience with class dictation has shown this to be time consuming, frustrating, and unreliable.

The author believes that dictation tests measure general proficiency in English, including many of the integrative skills in writing that they are easy to prepare; that they can be scored with consistency; and that they are much harder to cheat on than multiple-choice, completion, or cloze tests. On the negative side of the ledger, one needs to consider the fact that dictation tests are difficult to use for diagnostic purposes (they combine listening and writing); they are not usually helpful in measuring short-term progress; and they are not as easy to correct as multiple-choice, completion, or cloze tests. When constructing multiple-choice tests it is wise to keep in mind that there must be an empirical basis for selecting structure; it is the students who must supply the missing vocabulary word. Moreover, given the fact that a test is also a teaching instrument, it is fallacious to believe that interesting errors in a test will only serve the purpose of identifying knowledge in the learner. Errors, by the very fact of being encoded in the brain through reading or hearing, are counterproductive of learning except errors.

Fifthly, the test designer must bear in mind individual differences among

Conclusions

In conclusion, there are points that teachers responsible for selecting and/or developing TESOL courses and tests ought to consider.

Firstly, one should be aware that "as a general rule it is best to assess by a variety of test formats, the scores of which are taken as composite for reporting purposes" Weir (1988; 81). It may be a "cop-out", but, in the present state of knowledge, it is better than relying on one format with its biases and flaws.

Secondly, holistic scoring ought to be at least looked into. Brain child of Jacobs et al. (1981) who dichotomized frequency-count and holistic marking, said that it acknowledged the often vague and subjective approach of impression scoring and the restrictive quantification of analytic scoring. Cooper (1977; 5) defined holistic scoring as "any procedure which stops short of enumerating linguistic, rhetorical, or information feature of a piece of writing. " it is impressionistic, i.e. subjective. However, Jacobs et al. (1981) were of the opinion that "holistic evaluation by a human respondent gets us closer to what is essential in communication than frequency-count do". Somewhat reliable results can be achieved through the holistic method, if, as Chaplin (1970) recommended, achievement levels are clearly set and equated with grades.

Thirdly, goals must be written before embarking on any course or test design. Goals indicate the general performance, the global proficiency a student is expected to attain at the close of the course. These goals, broad and not necessarily measurable by definition, may best be assessed in behavioral terms and are the basis through holistic scoring. Next, objectives must be set. Objectives are best expressed in behavioral terms and are the basis for pre- and post-tests; they indicate the precise competencies the student is expected to attain at the close of the course. The teacher will then decide the relative weight to competency vs. the performance scores. If the purpose of the exercise for the student to pass traditional government examinations, then much greater weight should be given competencies as precisely quantified through tests. If the purpose of the exercise is to assist students to attain communicative proficiency in the foreign language, then much greater weight should be given to performances. A precise balance-fifty-fifty—is likely to indicate political compromise, pedagogical dilution, and personal faintheartedness.

Fourthly, the test-designer ought to decide early in his planning the types of tests which are likely to meet the demands of the goals and objectives. Oller (1987: 44) stressed the facts that the basic issue in testing is comprehensibility, in English (Overseas), and the Oxford Examination in English as a Foreign Language. Some of these tests are administered in consulates or other governmental agencies. An examination of published tests is a worthwhile exercise for any test-designer, because considerable time and money expenditure have gone into them.

In the United States, TOEFL has been the object of many studies, because it is the official admission instrument to most universities. It is therefore interesting to speculate on the test's predictive validity of academic performance. Hale, Stansfield, and Duran, (1984) found that academic performance on tests of English proficiency (primarily TOEFL) showed generally low correlation. This finding was in line with the fact that using TOEFL as a moderator variable along with admission tests has often yielded inconclusive results. As a matter of fact, other factors than TOEFL source were much more influential on predicting academic performance; for example, previous level of education (graduate students have the highest GPA mean, junior college transfers the lowest); geographical origins (Chinese students ranked among the highest scorers). Any conclusions from such studies can only be tentative, because various important parameters have not been included in the equations. For example, the Chinese students were highly selected group, were graduate students, and had political motivation. Data generally support the contention that there is some positive correlation between English proficiency and initial academic performance-correlation which fade rather quickly as students integrate and advance in their professional subject studies. Duran (1984) administered the MTELP test as well as the TOEFL to this experimental group and found that TOEFL is a well-structured test, with the students showing some relationship to each other but, at the same time, measuring somewhat different aspect of English knowledge. The correlation of 0.79 between the MTELP and TOEFL scores is consistent with earlier research showing a relatively strong relationship between these two measures of English proficiency.

Ilyin (1976) examined the confounding in measure of foreign language listening comprehension and hypothesized that the moderate – to – high relationship between measures of reading comprehension and listening comprehension on TOEFL and similar tests may be due to the structure of the typical test of listening comprehension, as it requires the subject to read the response alternatives. He concluded that the TOEFL Reading Comprehension subtest does not correlate more highly with the traditional Listening Comprehension subtest than with the experimental listening test. He thus came to a weak and tentative conclusion. He still does not know for sure how one test correlates with another, and why reading sometimes does and sometimes does not correlate with listening and speaking.

In the United States of America and in Canada, the Test of English as Foreign Language (TOEFL) is by far the most used for assessing the English language proficiency for admission to colleges and universities, as well as for placement by government and other agencies. It is composed of three separately timed sections using 4-choice objective question on listening comprehension (paragraph, short dialogue between two people, and passive comprehension), structure and written expression (completion, and error identification), and reading comprehension and vocabularies. Note that the listening comprehension test weekly tests for speaking ability—not enough, though, to constitute a valid test of speaking competence. The test takes approximately two hours. It addresses itself to adults who have complete secondary education (in America and Canada). Normally, it is not sold; rather the publisher, Educational Testing Service, administers the test on set dates at approved testing center. Nevertheless, there are a number of books written specifically for helping candidates take the test, and they can constitute a rich source of ideas and information for would-be testers. Moreover, the test is the administered abroad through some U.S. government agencies.

Another test popular in the United States is the Michigan Test of English Language Proficiency (MTELP). Although administered its popular (the University of Michigan) it is made available to schools. Its purpose and its target are the same as the TOEFL's. it consist of 100-items 3-part objective test, viz. 40 grammar items (sentence completion in 2-line dialogue, 40 vocabulary items, contextualized synonym and sentence completion), and 20 reading comprehension items based on 4 reading passages, each 100 or 350 words long. It takes 75 minutes for instruction- giving.

In Great Britain, there are many recognized examination organizations. One of the best known is British Council English Testing Unit which administers the English Language Testing Service Examination. It tests General Reading (i.e., sentence paragraph, cloze, comparison of 2 texts). Listening (visual identification, dialogue, appropriate, response, lecture), and Subject area English (i.e., reading, writing, and interview).

The Test of English (Overseas), sponsored by the Joint Matriculation Board (JMB) in Manchester, is a university entrance test composed of written English (3 pieces of connected writing, grammar/vocabulary, reading) and Aural English (dialogue, instructions, lecture, etc).

There are good many other tests for foreign students, e.g., (in the U.S.) the Comprehensive English Language Test (CELT), The Ilyin Oral Interview Test (IOI), the Basic English Skills Test (B.E.S.T), and the Interagency Language Roundtable Oral Interview (ILR). In England, among the many available, one finds the Communicative Use of English as a Foreign Language, the Test cloze score and oral interview score. This was true for both easy and difficult cloze passage. "The conclusions the researcher reached were that it appeared that "the predictive ability of various cloze test scores was dependent on whether speaking or writing was the criterion measure."

Cloze, of course, though currently fashionable, is not the only method of approaching competency measurement. Multiple-choice is a long-cherished method of teachers and testers alike, again if only because it seems to be an easy format to design. Yet, do cloze and multiple-choice measure the same aspects of, say, reading competency? No, according to Haskell (1976:15-16) who believed that a time cloze measured the process of reading, whereas multiple-choice measures the product of reading : the former looks at the reader's ability to understand the text as he is reading it, the latter verifies the reader's interpreting ability, i.e. the ability to abstract information for its meaning value.

A reader may comprehend the mutilated sentence as a whole and complete the pattern, but has he perceived and formulated the meaning of the text? Once more, if the form has been acquired (as measured by the cloze test), should the tester be satisfied that the reader has attained a communicative goal, i.e., meaningful communication between reader and writer? Or should he then administer a multiple-choice test to complete the evaluation? Much depends, of course, on the goals of the exercise: pass the State exam or learn to appreciate the foreign language as an information and cultural instrument.

Perhaps also the cloze test is not particularly indicated for measuring reading comprehension. It seems to do better in assessing syntax and lexis competence at the sentence level (Darnell 1968). Alderson (1978: 99) concurs that "cloze is essentially sentence bound."As such, then, cloze is not the ideal choice for assessing communicative skills. "

Off – the – Shelf Tests

It has become clear that, before designing tests, the tester must decide on what skills should be assessed: Reading comprehension? Listening comprehension? Writing? Speaking? On the assumption that a modern approach to EFL is concerned with the acquisition of all these skills, rather than in training students to take State examinations which are concerned essentially with what the student has learned about the foreign language, one might consider integrated tests rather than time-consuming separate tests for Each skill. The trouble is that there is not very much information concerning integrate tests. Consequently, we still have to rely on testing skills through separate instruments.

Johnson (1982:63) questions some assumptions about cloze testing. He believes that "the claims regarding the objectivity and automatic validity of cloze tests are largely false and because statistics are data and not arguments, valid conclusions can only be reached by processes of argument".

Johnson (1982) considers that the high levels of correspondence achieved in a number of studies involving cloze tests "may be regarded as resulting from and providing supporting evidence of, the reliability and validity of the judgment of the person who selected or prepared the texts rather than as evidence bearing upon cloze procedures per se."

This view is shared with Morrow (1979) who was suspicious of cloze and dictation as assessing instruments, because he felt they were testing competence rather than performance, knowledge about the language rather than knowledge of the language, i.e., proficiency in using the target language in authentic settings.

Mullen (1979:144) sought an alternative to the cloze test use in TESOL. He compared its performance with those of an editing test and with those of two direct tests of English proficiency (an oral interview and a writing task). He then investigated the relationships of these measures with TOEFL for subjects with TOEFL test scores. The editing test, based on a passage written at a seventh-grade reading level, required that the subjects cross out 50 words that did not belong in the passage. In the writing task, the subjects wrote an essay on a topic selected from a number of possible topics selected from a number of possible topics (No time limit was imposed). The correlation among the various measures indicated that "the nonidentification score on the editing test tended to correlate higher with the other measures than did the misidentification measure or a composite of the nonidentification and misidentification measures." Mullen, on the other hand (1979), investigated performance on a cloze reading test in ESL as a function of cloze passage difficulty and method of scoring. He also looked into the criterion of validity of cloze test performance in relation to performance on an oral interview, TOEFL, and a composite task.

Analysis of variance indicated that cloze test performance was most affected by individual differences among subjects and, to a much lesser degree, by method of cloze scoring, passage difficulty, and the order in which the passages were presented. Method of cloze passage scoring accounted for more variance in scores than did level of passage difficulty, despite the fact that the essay passage was rated at the seventh-grade level and the other passage was rated at the twelfth-grade level. Furthermore, Mullen reported that the correlation between exact-word cloze score and oral interview score "was not significantly different from the correlation between acceptable-word will affect the scores, and is a potential source of bias … and it is possible to complete items satisfactorily in the absence of any global understanding of the meaning of the text." After all, "what the receiver brings to the task of decoding is of far greater importance to eventual comprehension than the linguistic items on the page" (Hinofotis 1976:95). Indeed, the extra- linguistic data contained in the text or conversation and those brought by the learner may have much more interpretative force than the test items. Language is a many-way street: it interacts with space, time, and participants with their unfathomable personal worlds.

The Variety and Procedures of Tests

Some communicative-approach researchers, such as Oller (1973), are of the opinion that integrative tests do a rather good job of ascertaining performance in "real life" situations. They may not be ideal since they do not test the very acts of communicating in the target settings, yet they constitute practical and relatively reliable instruments. Other researchers such as Alderson (1978) doubt the validity of such tests. They believe that the cloze test, for example, is an unstable instrument, "because results can differ according to the starting-point and the rate of the deletions" Alderson (1978:225). He found that "individual cloze tests vary greatly as measures of EFL proficiency…. Changing the frequency of the test produces a different test, which appears to measure different abilities, unpredictably, Similarly, changing the text usually results in a different measure of EFL proficiency … {and} changes in scoring procedures also result in different validities of the cloze test, but the best validity correlation is achieved by the semantically acceptable procedure".

The cloze test is currently enjoying the greatest favor among test-making teachers. It appeals to them because it is rather simple to construct and it renders relatively credible data. Hirvala (1989: 9), however, remarks that the cloze test may be deceptive in its simplicity. "The undifferentiated use of the cloze procedure in a first and second classroom in hope of some kind of reading improvement is very dubious and is termed the "shotgun approach". Bartz (1979:91) stresses that its effective classroom implementation depends on careful text selection, preparation, and presentation.

The wavering in opinion in preference to one type of test rather than another is again subjected to the procedures of the selected test. That is to say, if a cloze test is our preference, then we have to ask ourselves: does the cloze test procedures measure the comprehension that is beyond the context /co-text? Although, as we have seen earlier, some believe that cloze test procedures produce tests that are generally consistent in the way they measure the language of the examinee, yet, there are some such as, Johnson (1982), who view cloze test procedures as inadequate in language testing.

The Purposes and Techniques of Testing

Tests are usually administered with a view to assessing competence, i.e., achievement or attainment of objectives. Yet, one of their most important functions is not to assess but to teach or to reinforce learnt knowledge. To a large extent, this is the function of daily quizzes, but it can fruitfully be extended to tests.

Green (1975:3-7) discusses the functions of tests and classifies them into three categories, viz., instructional tests, mastery tests, and measurement tests. Instructional tests are used in formative evaluations: essentially, they teach: mastery tests assess competencies, and measurement tests obtain norm-referenced or group-performance information (standardized tests): they give precise measurements of achievements. Saleemi (1988) prefers to label tests as evaluative, practical, instructional, and theoretical.

Any test should be reliable and valid. It should also discriminate, i.e. provide scores on the basis of which one can discriminate among members of a group. Evaluation can be norm-referenced or criterion-referenced. If the teacher or the educational system lays emphasis on group uniformity (as autocratic systems do), the traditional norm-referencing is then the way forward. If, on the other hand, the interest of the actors in the educational event is in the personal growth of the individual, then the student will be compared to himself or herself and to pre-set criteria as the educational process advances, i.e. one will be concerned with criterion-referenced testing, assessing to what degree the learner has reached the objectives of the educational or instructional process.

One warning worth heeding too is that "There is little if any reason to assume that conclusions from research with native speakers can validly be generalized to the case of non-native speakers" (Oller1973:112). This is one of the problems with EFL: much of the experience teachers and testers have had teaching English to native speakers is not transferable to EFL. Oller(1973: 112) reported that "with non-native speakers the method of allowing any contextually acceptable response is significantly superior to the exact word scoring technique . "He remarks that replacing words in the cloze or other like tests requires insights which may not be language skills at all. Doubtlessly, most such tests in along the same lines as do standard intelligence tests: they standardize on a restricted cultural and social population at a given time in history, and apply their logic to populations which have very different characteristics and operate in different socio-cultural contexts and at a later time in history.

Johnson (1982:63) further points out that "the intellectual content of a passage design a possible compromise on an analytical – synthetic basis and not if we believe that form and function are inextricably meshed.

Design-wise, a systems–approach to curriculum design will offer some advantages provided that the approach is adapted to the expected educational outcome rather than to the traditional instructional expectations. This means the administration of a pre- and post-test, so that achievement can be measured. As we have seen, achievement as demonstrated through competence in the testing situation is of aleatory value in terms of proficiency, i.e. adequate performance in the "real world" situation. For this reason, the pre- and post-test (an identical instrument) will have to be comprehensive and offer a mix of qualitative assessment as well as quantitative ones—not an easy undertaking.

In designing the format of all tests, the designer will heed Weir's admonition that "there is also evidence in the literature that the format of a task can unduly affect the performance of some candidates" (Weir 1988:83). Worse still, the very taking of a test may traumatically affect the performance of the testee who is afflicted with test anxiety. Does the student then get a Fail grade?

Although comprehensive, so-called integrative (such as summative) tests try to assess a number of linguistic skills, it remains to be seen to what extent they test communicative skills. Moller (1975: 5-10) noted that such tests do not require subjects to perform tasks considered to be relevant in the light of their known future use of the language. Morrow (1979:143) confirmed that tests such as dictation and cloze fail to provide opportunities for spontaneous production, that they measure competence rather than performance, and that the language norms are not the testee's but the examiner's. Oller (1979), on the other hand, believes that integrative tests, such as dictation and cloze, do have the capability to integrate disparate language skills as are applied in "real life" situations. Nobody doubts today that some linguistic competencies are necessary for communicative competence and performance in a foreign language by adults, but how linguistic competence relates to contextualized proficiency performance is still a mystery, and what tests best measure such performance is debatable . The major issue relating to the "communicative" approach to language testing is the generalizability of test results.

One decision the designer will have to take early is to what extent commercially produced tests should be adopted, and what tests (not quizzes, of course, which are the teacher's daily chore) should be made by the teacher rather than by the "expert". Before attempting to design tests, however, the teacher will do well to become fully familiar with the intricacies of test-design. By and large, teachers have had little schooling in this technical area.

foreign language proficiency. If one shares the communicative school's tenet that language cannot be dissociated from context and use, then the "emptiness of the audio-lingual approach" (Underhill 1982) will be recognized. If the pedagogical methodology allows individual students to develop at their own rate while "pushing" them to their i + 1 capacity, then validity is no longer a high-priority parameter. Unfortunately, a test can be highly reliable and yet have low validity, but it cannot be highly valid if it is not reliable. A dilemma for the test designer!

Yet, what if a test is both highly reliable and highly valid, but is impractical in terms of administration? It could be too cumbersome to administer if it is too expensive, difficult to score, or highly subjective in its interpretation. Clearly, the practically of the test has to be considered. The trend, at least in Western countries today, is methodological compromises based on political expediency and ignorance of how people learn and why and what they ought to learn.

Tests: How to choose and Use Them

The first question the teacher or test designer ought to ask herself or himself is: Who are the people I am supposed to teach or test? Of course, he or she could slant the question this way: Who are the people I am supposed to help learn the foreign language or to test? Indeed, "modern" education and instruction are learner-centered, and testing is situation centered. Methodologically, this makes a world of difference. Teachers no longer ask themselves: What do I think the students should learn and how should they learn and be tested? Rather, they ask themselves: Given the diversity of learning styles, of metacognitive backgrounds, of capabilities (talent?) in the skills of foreign language learning, of motivation, of the ability to negotiate meaning in a new social and functional context, and of ability to respond in testing situations, and given the availability of human, material, and time resources, how can I best help the individual students reach and verify the goals the educational system has set for them, perhaps willy-nilly but strictly enough that a general failure in reaching them will affect my pay-check? The first step, therefore, is to identify the target – population in terms of the goals – and possibly objectives – of course of study. The teacher could make profiles of individual students and of the class – an invaluable tool for the tester.

In terms of testing strategies, teachers and testers then ought to ask themselves: How can testing best predict the probable success of the students, given the resources at their and our disposal? Curriculum – wise, it has been found, for example, that grammatical competence "was not by itself a good predictor of communicative skills" as Savignon (1972:153-162) remarks. Does this mean that grammar can be excluded from the curriculum and the tests? Not if we performance .... In practice, a clear distinction between performance and competence will be difficult to maintain. As a working definition we might accept that communicative performance relates to the transmission and reception of particular meanings in particular contexts, and what can be tested is the quality and effectiveness of the performance observed under these circumstances." Weir (1988:81)

Reliability and Validity

Reliability and validity, the two principal criteria for evaluating tests, are indispensible. But given the disagreement on the objectivity of testing, these two criteria were also wavering in their importance in testing. For instance, in a discrete-point test beside the objectivity of the type of test, efforts were made to optimize reliability with only little attention paid to validity. While in the psycholinguistic- sociolinguistic approach, where communicative skills are under focus, reliability and validity received equal importance. The following is an assessment of the two criteria and their importance in language testing as seen by practitioners.

Underhill (1982:17-23), in an article titled The Great Reliability – Validity Trade–Off, remarks: "The two principal criteria for evaluating any kind of test are reliability (whether it gives consistent results) and validity (whether it measures what you think it does)." He notes that the main problem with tests of speaking and writing may simply be stated as: high reliability and high validity are seemingly incompatible. The situation is complicated by the existence of several different kinds of validity, some theoretical and intuitive and others empirical and quantifiable. As a result, what may be valid for one school of thought may not be for another. If you believe that real language use only occurs in creative communication between two or more parties with genuine reasons for communicating, then you may accept that the trade-off between reliability and validity is unavoidable.

Due to the artificiality of the testing situation, the question arises: How valid is testing? One of the problems in test design thus is the conflict between reliability and validity. Many researchers believe that quite often a certain degree of reliability needs to be sacrificed on the altar of validity. Weir (1988), for one, is of the opinion that indeed validity is the most important of the two parameters. Underhill (1982) states that for the audiolingualists "reliability was, and still is, considered to be logically prior to validity." Obviously, if one accepts the structuralists "discrete-point approach to testing", then reliability can be high, since the test reflects the items in the syllabus, rather than the events in the sociolinguistic situation. This is doubtlessly the better approach to instruction- to achievement testing, but may not be valid at all to education and to the learning and acquisition of to be complementary rather than incompatible. Thus it is possible to have an integrative test, or for that matter, a communicative one, that contains discrete-point items. The teacher should try to strike a balance between these two opposing tendencies. It seems judicious to attempt to construct tests that assess both form and use, that are both discrete-point and integrative.

Saleemi (1988) labeled the two main streams of pedagogical and testing approaches analytic and synthetic. The analytic approach is form-based, manipulative, atomistic, discrete-point based, and quantitative; whereas the synthetic is use-based, communicative, holistic, integrative, subjective, and qualitative. It is, of course, essentially the criteria of subjectivity and quality which bother test designers and school authorities, employers, and traditional parents who need to attach numbers to people to shed a glimmer on alleged intelligence and superficial achievements. But then, perhaps, test designers are at fault too, because so far they have not been able to understand the neurological, psychological, and social intricacies of language : they have not been able to design instruments to assess what, since Ferdinand de Saussure, the French, have called *le language*, i.e. the communication system which serves to communicate thoughts and feelings. It is a lot easier to measure *la langue*, the semiotic system which structures the transmission medium.

Rivera (1984), who provided a model of communicative competence that comprised linguistic and sociolinguistic dimensions, remarked that relatively little is known in fact about the communicative paradigm. Canale and Swain (1980:1-4) included in the concept of language competence grammatical competence (knowledge of the rules of grammar), sociolinguistic competence (knowledge of the rules of use and rules of discourse), and strategic competence (knowledge of verbal and non-verbal communication strategies). So, the communicative approach devotees increasingly feel that may be they have gone too far: that, after all, the media do affect the contents, that form not only follows function but shapes it through a feedback mechanism, in which case a compromise in testing for form as well as for function may be called for.

Traditionally, foreign-language testing has aimed at verifying competencies, i.e. measurable discrete-point achievements – which may not be compatible with performace in the intended setting. The conceptual error on the part of testers could generalize over tested achievements with ability to perform in the most probable settings: "Strictly speaking, a performance test is one which samples behavior in a single setting with no intention of generalizing beyond the setting –any other test is bound to concern itself with competence." Weir (1988:80). Competence, in this context, is construed as "knowing about the target-language" and can only be evaluated through its realization in

Oller (1987: 212), for example, negatively appraises the analytical method. His view is that discrete-point analysis necessarily breaks the elements of language apart and tries to teach (or test) them separately with little or no attention to the way those elements interact in a larger context of communication. What makes it ineffective as a basis for teaching or testing language is that the crucial properties of language are lost when its elements are separated. The fact is that in any system where the parts interact to produce properties and qualities that do not exist in the part separately, the whole is greater than the sum of its parts. Organizational constraints themselves become crucial properties of the system which simply cannot be found in the parts separately.

Spolsky (1985:189) praised the communicative approach and the coming of the psycholinguistic – sociolinguistic era that "was in many ways contrary to the allegedly atomistic assumptions of the 'discrete point ' test. "Furthermore, Read (1981:112) joined the ranks of current teachers of languages and noted that "from a psycholinguistic perspective, language came to be seen as less of a well-defined taxonomic structure and more of a dynamic, creative, functional system."

The sociolinguistic contribution centres on the concept of communicative competence. "Clearly, however, there is no method without disadvantages and such an entangled conception no doubts complicates testing. Here are only some of the perplexing questions: How can one grade a "dynamic, creative, functional" system? Against what standards? With what instruments?

There is a difficulty of predicting language learning output on the basis of instructional input. Weir (1988 : 5-10) believes that "the difference between knowing how to analyze input and knowing how to construct output would seem to outweigh the correspondence between the two processes … Correlational data do not provide evidence about standards."

As far as Oller (1973: 105-18) is concerned, the more one contextualizes language (a sine qua non of communicative teaching), the better language is perceived, processed, and acquired. Again, communicative contextualization broadens the scope of education, but makes it harder to relate instruction to educational objectives and to testing linguistic proficiency in the social and functional settings of the target- language. It is easy to test for competency – based behavioral objectives, because one is dealing with a closed system. No empirical research has clearly established the relationship between the two concepts and methodologies.

Among those who believe in an eclectic method of testing is Saleemi (1988: 2-6), who thinks that the analytic and synthetic approaches are by no means mutually exclusive. Most of these polarized concepts are likely to prove

The term integrative test was used in the 1960s and was intended to have "holistic" test procedures, which include oral interview, composition and dictation. The interest in this approach came as a reaction towards another type of test, which was dominating the field and was called the discrete–point test. This type of test can be defined as the one in which language features or items were tested in isolation and therefore the approach which advocates this type of test is called the analytical approach and the method was refered to as the psychometric.

After the explanation of terms, an analysis of the use of these terms will be provided for further explication of the variation in terms of language testing. Notwithstanding that applied linguists and language instructors / practitioners during their assessment of testing and its techniques were motivated by different objectives. Therefore, their ways of looking at tests are subjective to some extent. Their advocation of one type of method of testing rather than another was directed by their specific needs and purposes. Hence the names that they gave to different methods were indeed relative. For instance, Spolkey's identification of the phases of language testing was instigated, in the first place, by his interest in the description and measurement of problems that were very delicate and quite tenuous in linguistics and psychological investigation where decision is the first goal .That is to say, if learning a language is a skill, this skill has to be tested on the examinee's performance of the language, especially if examination or testing was to provide qualification.

On the other hand, the psycholinguistic approach was a preference to other language instructors/ practicioners because this type of test was the ideal way for showing success in their methods of teaching and hence testing efficiency as a criterion measurement was very essential for their mechanism of teaching. That is, assessing the leaner's ability or capacity to use language components or items integratively by integrating different aspects of skills.

The case in point here is that variation in terminologies as well as in approaches to language testing was determined by their limitations in relation to the criteria of validity, reliability and efficiency in language testing.

Yet there are those who believe that both approaches, i.e.., analytical and integrative can be combined and their techniques in testing can also be integrated. Davies (1978: 127-59), for instance, noted a tension between the analytical approach to testing and integrative one. He argues that language testing should be based on a combination of both views. In any case, he comments, no test could reasonably be wholly analytical or integrative. The following review of approaches to language testing will be sufficient to show how advocates of certain approaches treat their methods as "distinct" or "pure" as Weir (1988) has described them.

The Problems and the Background

Tests in instruction and education have been said to be used with a view to assessing "the effectiveness, efficiency, and suitability of the materials in relation to the instructional objectives" (Romiszowski 1986: 401). As simple as this statement seems, it demands a lot of education. One of the difficulties facing all players in the testing game is that "language and language ability are abstract theoretical entities" (Oller 1987:44).

Because of this abstractness and the high complexity of language, linguistic theories have flourished and withered often to resurface in modified form not only as our knowledge of linguistic processes has increased but as our philosophical, social, and political trends have dictated. To go back to the mid 1970s only, tests in instruction and education have developed from discrete-item tests to integrative tests, from the viewpoints of structural linguists through those of psycholinguists to those of sociolinguists and admirers of communicative approach .

The change in emphasis in language teaching resulted in change in language testing. Yet, the techniques and theories of language testing were rather insusceptible to change in procedures. Nevertheless, the growing interest in and concern with the social and psychological dimensions of language use had forced the advocates of this approach to develop and evaluate their testing instruments.

It is a well known fact that life is full of testing, but why and how we test are still elusive. This article aims to review how language tests are observed from the point of view of educators and language instructors.

The article will be divided into two main sections. Firstly, taking into consideration the pervasive range of terminologies, it is imperative to explain the major concepts related to the subject–matter examined in this article. Secondly, an assessment of the implications of these concepts will be addressed through the viewpoints of the specialists and practitioners in the field.

There is a fair amount of terminologies that deal with and describe different aspects of testing. In the following a brief definition will be given of these terms that have direct relation to the issues to be discussed in the work.

Abstract

An Analytical Study of Testing Effectiveness

Dr. Khalid AlKhaja
Dr. Maryam Baishak

The testing and assessment of instruction and education are supposed to measure the effectiveness and adequacy of teaching/learning resources (material and human) in relation to the objectives of the instructional or educational system.

Language is a highly abstract and complex communication system and does not lend itself to easy analysis and testing. As a result, language theories have been unable to propose universal pedagogical methodologies. This article aims to identify some of the theoretical and operational problems of TESOL and, to the extent possible; it strives to present recommendations for teachers and test- designers. It confronts the two major streams of approaches to testing: The analytical and the integrative. The integrative –or synthetic-approach stresses communicative skills as opposed to discrete-point fragmentation of the communicative event. Sociolinguistic and learner-centered communicative approach, thought laudable in its intent, has failed to develop testing instruments with acceptable reliability. Today, many researchers combine the two approaches, because proficiency is hard to achieve without competency. One of the problems facing the test-designer, in this perspective, is what weight to ascribe to each approach. The apparent antagonism between quantification and qualification input might well be impossible to assess. The communicative approach sees contextualization as the key to improved language perception and processing.

An Analytical Study of Testing Effectiveness

Dr. Khalid AlKhaja
Dr. Maryam Baishak

* Assistant Professor in Applied Linguistics
Dean, College of Information, Mass Communication and Humanities - Ajman University of Science and Technology.

* Assistant Professor in Linguistics
Assistant Dean for students' achievements, United Arab Emirates University

UNITED ARAB EMIRATES-DUBAI
COLLEGE OF ISLAMIC & ARABIC STUDIES

ACADEMIC REFEREED JOURNAL OF
ISLAMIC & ARABIC STUDIES COLLEGE

ISSUE NO. 37
Jumada 2, 1430H - June 2009CE

ISSN 1607- 209X

This Journal is listed in the **"Ulrich's International Periodicals Directory"**
under record No. **157016**

e-mail: iascm@emirates.net.ae

UNITED ARAB EMIRATES-DUBAI

COLLEGE OF ISLAMIC & ARABIC STUDIES

Islamic & Arabic Studies College Magazine

Academic refereed journal

Read In This Issue

Issue No. 37

E Mail iascm@emirates.net.ae

Website www.islamic-college.co.ae

1430هـ/ 2009م