DOCUMENT RESUME

ED 365 149 FL 021 765

AUTHOR Milanovic, Michael
TITLE Materials-Based Tests: How Well Do They Work?
PUB DATE 91
NOTE 21p.; In: Sarinee, Anivan, Ed. Current Developments in Language Testing. Anthology Series 25. Paper presented at the Regional Language Centre Seminar on Language Testing and Language Programme Evaluation (April 9-12, 1990); see FL 021 757.
PUB TYPE Speeches/Conference Papers (150) -- Viewpoints (Opinion/Position Papers, Essays, etc.) (120)

EDRS PRICE MF01/PC01 Plus P.stage.
DESCRIPTORS *English (Second Language); *Instructional Materials; *Language Tests; Second Languages; *Teacher Developed Materials; Teacher Education; Teacher Role; Test Construction; *Test Content; *Testing

ABSTRACT

It is argued that tests of English as a Second Language based on specific instructional materials, reflecting both real-world and classroom language activities, can satisfy measurement needs and provide interesting psycholinguistic insights. It is also proposed that the potential for mismatch between test and materials can be overcome by several strategies, including: integrating testing programs into institutional life, rather than treating testing as ancillary to learning; using test content that reflects classroom activities and student lives; and active encouragement of teachers' involvement in test writing, including ap. opriate training. (MSE)

MATERIALS-BASED TESTS: HOW WELL DO THEY WORK?

Michael Milanovic

INTRODUCTION

While all language tests tend to be materials-generating, their rationale and format is varied and they have differing effects on classroom practice. I would like to propose that language tests can be described as measurement-based, psycholinguistically-based and materials-based. Measurement-based tests tend to use a restricted item format, most commonly multiple-choice. They claim high reliability though are often criticized for lack of face and content validity. Psycholinguistically-based tests also tend to use a restricted range of item formats, such as cloze and dictation. It has been claimed that such tests tap an underlying language competence but they too have been criticized for lack of face and content validity. Materials-based tests arise out of trends in the development of language teaching materials. In recent years the most dominant generator of materials-based tests, in the British context at least has been the communicative language teaching movement. One important feature of materials-based tests is their use of a wide range of item formats which attempt to reflect teaching materials and currently, real-world language performance. Communicatively generated materials-based tests have tended to stress face and content validity but have placed less emphasis on reliability.

Materials-based test construction tends to be dynamic. New item formats are developed in line with developments in teaching methodology and materials. Measurement and psycholinguistically-based tests, on the other hand, tend to be static. The range of item formats does not change dramatically.

It is important to note that the distinctions made above are not clear cut. An item format may be materials-based when it is first developed in that it represents current trends in teaching methodology or views of the nature of language competence. If it then becomes established, and continues to be used, despite changes in methodology or views of language, it is no longer materials-based. Ideally, tests should be materials-based, psycholinguistically-based and measurement-based concurrently. Only when this is the case, can we claim to have reliable and valid tests.

Hamp-Lyons (1989) distinguishes between two types of language testing research. The first is for the purposes of validating tests that will be operationally used. The second, which she calls metatesting, she defines as having its purpose in:

> "... *the investigation of how, why and when language is acquired or learned, not acquired or not learned, the ways and contexts in which, and the purposes for which, it is used and stored, and other such psycholinguistic questions*".

This type of language testing research has focused to a great extent on psycholinguistically and measurement-based test types and less on materials-based ones. In so doing, it has laid itself open to the criticism that too much attention has been paid to too restricted a range of item types. That not enough attention has been paid to understanding the interaction between background variables such as proficiency levels (Farhady, 1982) or the effects of the learning/teaching environment (Cziko, 1984) on test performance. The same might be said with regard to a systematic description of test content and the interaction between content and performance. Serious interest in this area is relatively recent (Bachman et al. 1988).

The aim of this article is to show that materials-based tests of English as a Second/Foreign language, reflecting both real-world and classroom language activities, can satisfy both measurement demands and provide interesting psycholinguistic insights. In other words, that there need not be an overpowering tension between the three perspectives outlined above. In practical terms, the tests and procedures used as examples here are most directly relevant in the context of a language teaching institute.

Test constructors, educators and test consumers need to be satisfied that tests are measuring what they are intended to measure consistently and fairly. Tests must be reliable because people's lives may depends on the results. For a variety of reasons it appears to be the case that many test construction agencies have been too willing to believe that satisfactory measurement criteria can only be achieved in a limited number of ways. In language testing, although this is also true in many other subject areas, this belief has led to the development and very wide use of indirect methods of testing ability. The most common such method is the multiple-choice item. It satisfies the conditions of objectivity of marking, economy of scoring and readily lends itself to statistical validation procedures. However, it does not have a very good effect on classroom practice, nor does it reflect the way language is used in real-world contexts.

A tension exists in the language teaching/testing world between the need for accountability in the educational process and the need to be accountable for the effects of testing on the educational process. In other words, while we must be able to trust the testing instruments that we use, it must be accepted that tests have a major influence on what goes on in the classroom. Both teachers and students generally believe, and rightly so to a great extent, that one of the best ways to prepare for a test is to practice the items in the test. It is a well established fact that the multiple-choice test format does not inspire innovative methodology, that it has had a largely negative effect on classrooms all over the world. Unhappily, it is still widely considered the best testing has to offer because it satisfies the need for measurement accountability and is economical to administer and mark.

In test validation research the problem of relating testing materials to useful and beneficial teaching materials has led to investigations of different test formats. Swain (1985) describes a Canadian project in which students actually participate in the creation of test items based on their own perceived needs. Swain formulates four principles that should guide the test constructor. These are:

i start from somewhere;
ii concentrate on content;
iii bias for best;
iv work for washback.

The first principle, start from somewhere, suggests that the test constructor needs to base test development on a model of language ability. The second principle, concentrate on content, suggests that test content should motivate, be substantive and partially new, that it should be integrated, and that it should be interactive. The third principle, bias for best, demands that tests should aim to get the best out of students, rather than the worst. Swain feels that it is important to try and make the testing experience less threatening and potentially harmful. The fourth principle, work for washback, requires that test writers should not forget that test content has a major effect on classroom, practice and that they should work towards making that effect as positive as possible. Clearly, these four principles cannot be satisfied by using only indirect measures such as multiple-choice items. We have to turn towards other item types.

There have been attempts originating from testing agencies to make language tests more relevant and meaningful. The Royal Society of Arts (RSA) in the United Kingdom developed a series of examinations in the Communicative use of English in the late seventies based on criteria proposed by Morrow (1979). The tasks appearing in these examinations attempted to reflect authentic communication activities and current trends in language teaching methodology. Great emphasis was placed on the involvement of language teachers in test construction and marking, and the backwash effect of this process, as well as the examinations themselves, on the teaching of English. It must be said that these are powerful features of the approach taken by examining boards in Britain. Examinations are not perceived as the property of boards alone. Ownership is distributed between the boards, methodologists and teachers, all of whom accept responsibility for the effect that the examinations have on the consumer - the students taking examinations - and the educational process. Many examining boards in the United Kingdom try to reflect language in use in many of the item types they use. This has been done in response to pressure from teachers demanding an approach that reflects more closely recent trends in methodology. The trend towards more realistic test items has not always been backed up by the equally important need to validate such tests. The combination of innovation and appropriate validation procedures is a challenge yet to be fully faced.

Even so, the examples cited above show that parts of the testing world are trying to move towards tests that look more valid and try to reflect both real life language activities and recent trends in language teaching methodology and materials more closely.

A major strength of the materials-based approach is that it actively works for positive washback effect. This helps to indicate to students, as well as teachers, that the main purpose of language instruction is to prepare students for the world outside the classroom. This should give the materials-based approach significant motivational value. However, as Wesche (1987) points out with regard to performance-based test construction (and the same is surely true with regard to materials-based tests):

> *"Performance-based test construction requires considerable advance or 'front end' work: careful specification of objectives, identification and sampling of appropriate discourse types, content and tasks, and consideration of scoring criteria and procedures."*

When preparing materials-based tests, achieving reliability may appear to be difficult due in part to the untried nature of many of the item types and in part to the fact that achieving reliable measurement is always a problem. However, both reliability and validity have to be established. Extensive investigation, moderation and pretesting procedures have to be employed to achieve both reliability and validity at the expense of neither.

While several attempts have been made to produce face, and to some extent content valid language tests, a disturbing lack of attention has been paid to making such tests reliable, or establishing their construct validity. In the following I will describe a project that attempted to produce a test battery that was based, to some extent at least, on the real world needs of the test takers. It took place in the British Council language teaching institute in Hong Kong.

The British Council language institute in Hong Kong is the largest of its kind in the world. There are between 9,000 and 12,000 students registered in any one term. In the region of 80% of the students are registered in what are loosely called General English courses. In fact this term is misleading. Through a fairly standard ESP type of investigation into the language needs of the students, it was possible to show that two main categories of student were attending courses. These were low to middle grade office workers, and skilled manual workers. This meant that the courses could be designed with these two main categories in mind. A much smaller third category was also identified, though this overlapped heavily with the first two. This category was students learning English for varied reasons. A set of real-world English language performance language performance descriptions were generated. These formed the basis for test specifications and the generation of teaching materials.

TEST CONTENT

An achievement or progress test should reflect course content. This is not to say that each item in the course needs to be tested. Unfortunately, in the minds of many teachers and students a test needs to cover all aspects of a course to be valid or fair. If the test is a discrete-point grammar test, testing a discrete-point grammar course then this may be possible if not desirable (Carroll, 1961). In almost any other context it is simply not possible to test all that has been taught in the time available for testing. In deciding test content the following points need to be considered:

i A representative sample of areas covered in the course need to appear in the test. (the term 'representative' is not defined accurately. Its meaning will vary from context to context, and test to test);

ii Enough variety needs to be present to satisfy teachers and students that no one is being discriminated against or favoured in any way.

iii The item types that appear in a test must be familiar to both teachers and students.

iv The test content must not appear to be trivial.

v There must not be an undue emphasis in the test areas of minor importance.

vi The use of item formats suited primarily to testing purposes eg. discrete-point multiple-choice, should be avoided as far as possible if they conflict with sound teaching principles (whatever these may be).

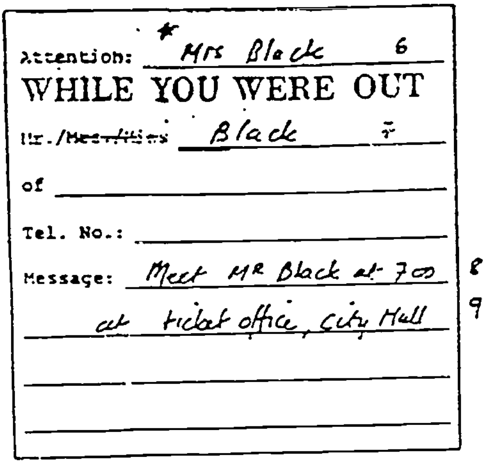
All too often operationally used tests do not resemble teaching materials in style and format. If, teaching a language aims to prepare learners for real-world use of that language then it is reasonable to assume that certain tasks encountered in the classroom will, to some extent, reflect reality. Other tasks may be of a purely pedagogical nature. There must, for students and teachers, be either a pedagogical or real-world familiarity with items in a test - preferably both.

Items to be included in tests should be selected on the basis of their relevance and familiarity and the extent to which they are, when incorporated into a test, reflective of the course students followed and the ways in which they put language to use.

TASK-BASED VS DISCRETE-POINT ITEMS

The argument above raises the question of whether test items should be task-based or discrete-point. As teaching becomes more whole-task-based it is inevitable that test items must follow. However, this causes two sets of problems from a testing point of view. Firstly, how is the tester to sample effectively from all the task-based activities and to what extent are the results obtained generalizable? These problems have been discussed at length over the years but no satisfactory solution has been reached.

Secondly, in real life, a task is generally either successfully completed or not. In class, the teacher can focus on any aspect of the task in order to improve student performance. In the testing context, however, the task may provide only one mark if treated as a unity, as long as an overall criterion for success can be defined and whether this is possible is a moot point. Such a task may take several minutes or longer to complete. If the test in which it resides is to be used for ranking or grading it can be extremely uneconomical to treat a task as a single unit. An example of a task based item would be the telephone message form illustrated below.



Attention: *Mrs Black 6

WHILE YOU WERE OUT

Mr./~~Mrs./Miss~~ Black 7

of

Tel. No.:

Message: Meet MR Black at 700 8

at ticket office, City Hall 9

6.	1	0	0
7.	1	0	0
8.	1	0	0
9.	1	0	0

Clearly, for the task to have been successfully completed all the relevant information needs to be present. Unfortunately this is rarely the case - mistakes are made, information is missing. It would be difficult to score such an item dichotomously and achieve a reasonable distribution of scores or provide enough information for effective test validation.

A compromise solution that satisfies the criterion of authentic/realistic appearance, allows the tester to allocate an appropriate number of points to the task to make it economical from a scoring point of view, and provides relevant data for validation, is to break a task down into discrete points for marking purposes. It is important the student does not perceive such a task as a group of individual items but rather as a whole task.

CONSULTATION IN TEST CONSTRUCTION

The views of both students and teachers are important in test construction. It is difficult to involve students in test construction, but it is of great importance that their views are sought after pre-testing or test administration in order that objectionable items can at least be considered again. It is often enough for teachers to ask for informal feedback at the end of a test. Some recent research has also focused on introspection by students.

Equally important as the views of the students is that of the teachers. At best the concept of testing in English language teaching is unpopular and badly understood. For any approach to testing to succeed, therefore, three factors are of vital importance:

i Teachers must gain some familiarity with the principles and practice of language testing. This is perhaps best achieved through some form of basic training course;

ii Teachers must be involved in the process of test design, item format selection, and the writing of test items;

iii Teachers must be familiar with the life cycle of a test and aware of the fact that good test construction cannot be haphazard.

It is unfortunately very difficult to achieve any of the three aims in a short period of time with an entire teaching body of any size. In the case of the British Council institute in Hong Kong, there were more than one hundred teachers employed at any one time and so, training and involvement had to take place by degree. However, it was anticipated that the credibility of the tests and the process of consultation would be better accepted when those who were actually involved in working on the tests mixed with teachers who were not involved. The more teachers could be made to feel a personal commitment to the tests, the more people there were who would be available to explain and defend them as necessary. The image of the test constructor in the ivory tower having no contact with the teaching body had to be dispelled as fully as possible. Thus it was that there were generally between four and six teachers involved in test construction in any one term.

A MATERIALS-BASED TEST

One of the tests in the battery developed in Hong Kong will now be described in order to illustrate some of the points made earlier. The A3 Progress test, like all the others, is divided into four basic parts. A3 level students have a fairly low standard of English therefore the test tasks they have to perform are of a rather basic kind. Every attempt was made, however, to keep these tasks realistic and relevant.

The Listening Test, a copy of which appears in appendix 1, comprises three item types. The first simulates a typical telephone situation that the students are likely to encounter, the second a face to face exchange at a hotel reception desk, and the third a face to face exchange between a travel agency clerk and a tourist booking a day, tour. The skills tested are listed below:

Taking telephone messages

This involves:

- writing down spelling of names;
- writing down telephone numbers;
- writing down short messages (instructions, places, times).

Writing down information about a customer

This involves:

- writing down spelling of last time;
- writing down first name when not spelt;
- writing down 'Tokyo' (not spelt);
- writing down spelling of address;
- writing down name of local airline (not spelt).

Writing down information for customers at a travel desk

This involves:

- writing down spelling of name;
- writing down room number;
- writing down number of people going on trip;
- writing down times of day;
- writing down price.

In the real world, skills frequently tend to integrate. This feature of language use was accepted as fundamental to item design. However, it should be noted that reading and writing are kept to a minimum in the Listening test. It was felt that it would be unfair to include a significant element of either of these two skills, since the students' competence in both was likely to affect performance in listening. Enough reading and writing was retained to ensure the reality of the tasks while not hindering students in their completion of these tasks. The tape recordings were made in studio conditions and various sound effects incorporated to make them more realistic.

The Grammar Test caused some concern. It was decided that the tests should include a section on grammar, or perhaps more appropriately, accuracy. The communicative approach has been much criticized by teachers and students for its perceived lack of concern for the formal features of language. In the Hong Kong context, it was very important to the students that there should be something called grammar in the tests. From the theoretical point of view, it was also felt that emphasis should be placed on more formal features of language. How they should be tested was the difficult question. If standard discrete-point multiple-choice items were used, the washback effect on the classroom would have been negative in the sense that the multiple-choice approach to grammar teaching was not a feature of the teaching method in the British Council. It was also thought better to use an item type which was text-based as opposed to sentence-based. To this end a variation on the cloze procedure was developed for use in the lower level progress tests. It was given the name 'banked cloze' because, above each text, there was a bank of words, normally two or three more than there were spaces in the text. Students chose a word from the bank to match one of the spaces. Each text was based on some authentic text-type relevant to and within the experience of the students. These were:

An article from Student News.
A newspaper article.
A description of an office layout.
A letter to a friend.

It should be pointed out that the same format was not used at higher levels. A method of rational deletion (Alderson, 1983) was used instead. It was accepted that there were many potential hazards in the use of the cloze. However, it satisfied the washback requirements better than any other item-type available at the time.

The Appropriacy Test was based on the common teaching technique, the half and half dialogue. Situations relevant to and within the experience of the students were selected. One person's part of the dialogue was left blank and it was up to the student to complete it as best he could. Clearly, writing down what would be said in a conversational context suffers from the point of view that it is not very realistic. However, it was a teaching device commonly used in the institute, and thus familiar to the students. Furthermore, it focused attention on the sociolinguistic aspects of language and allowed for a degree of controlled creativity on the part of the student. The marking was carried out on two levels. If the response was inappropriate it received no marks, regardless of accuracy.

If it was appropriate, then the marks were scaled according to accuracy. Only a response that was both appropriate and wholly accurate could receive full marks.

The types of functional responses that the students were expected to make are listed below:

- giving directions;
- asking about well being;
- offering a drink;
- asking for preference;
- asking about type of work/job;
- asking about starting time;
- asking about finishing time;
- giving information about own job;
- giving information about week-end activities.

Reading and Writing were the final two skills areas in this test. An attempt was made here to integrate the activity as much as possible, and to base the task on realistic texts. Students were asked to fill in a visa application form using a letter and passport as sources of information. The passport was authentic reading material, while the letter was especially written for the test. The form was a slightly modified version of a real visa application form. The introduction of authentic materials into the test as opposed to contrived teaching materials, and a focus on a situation that any of the students may need to deal with was an important statement. The test was attempting to do something that, at the time, most of the teachers were not, that is, using authentic materials with low proficiency students. The teachers soon saw that the nature of the task was as important as the material. They were also able to see that students almost enjoyed this sort of activity, and immediately understood its relevance to their day-to-day lives. Informal feedback from teachers, after the introduction of the test, indicated that it had encouraged a greater focus on the use of authentic materials and realistic tasks in the classroom. It seemed that positive washback was being achieved.

THE TEST CONSTRUCTION PROCESS

Little guidance has appeared on how to actually develop a communicative test battery or integrate it into the workings of a school environment. Carroll (1978; 1980) gives the matter of test development some coverage but he does not consider, in any depth, the consequences or role of testing in an educational context. With regard to involving teachers and integrating testing into the school environment, there is also very little guidance available. Alderson and Walters (1983) discuss the question of training teachers in testing techniques on a postgraduate course. The process of training and sensitization in-service is not considered.

Inextricably linked to the process of test development, as described here, is the need to actively involve and train teachers in the institute in test design and implementation. The tests developed in Hong Kong underwent very similar treatment before they were finally implemented. It was through involving teachers in the stages of this treatment, that some degree of training and sensitization was achieved. Listed below are the six stages of test preparation. I believe they are appropriate to many situations where teaching and testing interact.

Stage 1

Test construction needs to be coordinated. At the beginning of a test construction cycle, the testing coordinator needs to meet with a group of test item writers, normally teachers, specializing in writing items for a given test. In this case 'specializing' means teachers who have worked with students at a given level and are preferably teaching them. The purpose of a preliminary meeting is to discuss any ideas that the teachers may have, to take into account any feedback regarding the tests already operating and decide on a topic area that each teacher could focus on in order to prepare items for the next meeting. Teachers need to be briefed on some of the difficulties they are likely to encounter in test item writing, and how they might cope with such difficulties.

Stage 2

The teachers write first draft items in light of Stage 1 discussions, their experience of the materials and students, the course outlines and performance objectives.

Stage 3

A series of meeting is held when the items prepared by individual teachers are subjected to group moderation. The items are discussed in terms of their relevance, testing points, importance, and suitability for the students in question. It is important that any idiosyncrasies are removed at this stage.

Group moderation is a vital phase in the preparation of items for several reasons. Firstly, in test construction, where great precision and clarity are required, several people working on an item inevitably produce better results than just one person working alone. Secondly, a group product is generally better balanced and more widely applicable if worked on by teachers all actively engaged in teaching a course. Thirdly, the teachers in the test construction team are well prepared for many of the questions that might later arise from the use of a particular item and are able to justify its inclusion in a test.

Teachers are often found to rush moderation at first because they may be worried about offending their colleagues or unable to focus precisely enough on the likely problems or difficulties an item may pose, such as markability, reasonable restriction of possible answers and so forth. It is important to insist on thorough moderation at this stage since without it the product will probably be of inferior quality and may need complete re-writing and pretesting before it is of any use.

Stage 4

Completed items are then informally trialled with participating teachers' classes in order to uncover any glaring difficulties that the moderation team had not been able to predict. This helps to greatly increase the sensitivity of teachers engaged in item writing. It is all too commonly believed by teachers and administraters alike that test construction can be accomplished quickly and that the product will still be quite acceptable. Unfortunately, due to a number of factors such as the unpredictability of the students, the shortsightedness of the test writer, the lack of clarity in instructions, this is rarely the case. Initial moderation helps to make teachers aware of some of the difficulties; trialling informally with their own classes is an invaluable addition to this sensitization process. Moreover, teachers have the opportunity of observing the reactions of students to the items and the way in which they attempt to do them. Both of these factors are very important in the construction of task-based tests that attempt to have a positive washback effect on the classroom.

Enough time needs to be allocated to Stages 1-4. In the context of a teaching institution, given the range of demands on everyone's time, at least three or four months is required for the successful completion of these stages.

Stages 5

After initial trialling, the moderation team meets again, and in light of the experience gained so far prepares a pretest version of a test or part of a test.

The pre-test is then administered to a representative sample of the population and the results analyzed. It is generally necessary to pre-test up to twice as many items as will eventually be required to achieve the appropriate quality.

Stages 6

The moderation team meets to discuss the results of the pretest and decide on the final form of the test items.

Any test item generally takes at least six months from inception to completion in the context under discussion here. Teachers should be involved in the process from start to finish. Those teachers involved realize that the process of test construction, while lengthy and time consuming, must be carried out with the greatest of care because the test results have a very real influence on the students in question. They are able to bear witness to the fact that no test can be produced without due care and attention. To begin with, most of them believe the approach to be unnecessarily long drawn out and tedious, but as they work on items and become fully aware of the fallibility of tests and test constructors, their attitudes change.

Do these tests meet measurement criteria?

I made the claim earlier that materials-based tests need to function at least as well as measurement-based tests, from a statistical point of view. Even if the same degree of economy of marking cannot be achieved, this is out weighed, in an institutional context, by the considerable educational benefits.

Some basic test statistics for five progress tests from the battery in question are presented below. Each test was analyzed in two ways. Firstly, it was treated as a unity, in the sense that none of the sections were analyzed separately. This means that the mean, standard deviation, reliability and standard error of measurement were established for the whole test. Then each section was treated as a separate test. This meant that there were four separate analyses of Listening, Grammar, Appropriacy, and Reading and Writing.

Table 1.

	WT	LIS	GRM	APP	RD/WT
A3 Test					
$\bar{X}$	63%	55%	60%	81%	69%
SD	19%	24%	22%	24%	28%
KR20	0.95	0.92	0.88	0.84	0.92
NQ	89	28	29	10	22
NS	264	264	264	264	264
B1 Test					
$\bar{X}$	54%	42%	52%	77%	53%
SD	16%	20%	21%	18%	28%
KR20	0.93	0.87	0.83	0.78	0.89
NQ	96	33	24	19	20
NS	305	305	305	305	305
B2 Test					
$\bar{X}$	58%	42%	57%	74%	65%
SD	14%	18%	18%	15%	19%
KR20	0.91	0.82	0.80	0.68	0.85
NQ	99	29	24	20	26
NS	259	259	259	259	259
C1 Test					
$\bar{X}$	57%	55%	46%	80%	64%
SD	16%	20%	19%	23%	24%
KR20	0.94	0.88	0.86	0.84	0.91
NQ	112	34	35	12	31
NS	250	250	250	250	250
C2 Test					
$\bar{X}$	58%	57%	49%	79%	62%
SD	18%	20%	21%	22%	27%
KR20	0.95	0.86	0.87	0.74	0.91
NQ	98	31	31	09	25
NS	242	242	242	242	242

KEY

WT	=	Whole Test
LIS	=	Listening
GRM	=	Grammar
APP	=	Appropriacy
RD/WT	=	Reading and writing
X	=	mean score;
SD	=	standard deviation
KR20	=	Kuder-Richardson 20 reliability quotient;
NO	=	number of items in the test or subtest;
NS	=	number of students in the sample

Table 1 illustrates basic overall test and subtest statistical characteristics.

It is clear from these figures that the tests are very reliable. The reasons for this are as follows:

i. much time and effort was put into planning and moderation;

ii. test content was relevant and well defined;

iii. teachers were involved in the process of test writing from the earliest stages;

iv. the tests were all pretested and revised in light of pretest performance.

Do these tests meet psycholinguistic criteria?

Meeting psycholinguistic demands is a complex issue at several levels. In this context, the most straightforward of these is to attempt to show that the subtests are indeed measuring different aspects of underlying language performance. In order to do this it is necessary to demonstrate that tasks in a subtest relate to each other more closely than they do to tasks in other subtests. The most widely used methodology to investigate this type of issue is factor analysis. Simply put, factor analysis is a correlational technique which attempts to reduce the number of observed variables to a smaller number of underlying variables. It does this by grouping the observed variables on the basis of how closely related they are to each other. It is then up to the researcher to interpret the findings.

In the case of the tests in the battery described here this was done by computing students' scores on subtest tasks and then treating these tasks as mini-tests in their own right. If the tasks grouped together according to the skills they were said to be testing, then this would provide evidence that performance could be accounted for by different underlying skills. A factor analysis for the A3 test is illustrated in Table 2.

Table 2.

Subtest	Factor 1	Factor 2	Factor 3	Factor 4
Listening 4	.74541			
Listening 1	.70287			
Listening 5	.64940			
Listening 2	.64851			
Listening 6	.63182			
Listening 3	.62097			
Grammar 1		.75096		
Grammar 4		.69953		
Grammar 2		.63289		
Grammar 3		.51338		
Approp 1				
Rd/Wrt 4			.86169	
Rd/Wrt 2			.65637	
Rd/Wrt 5			.59547	
Rd/Wrt 3		.41049	.52075	
Rd/Wrt 1			.44136	
Approp 2				.75125
Approp 3		.41395		.54720

Interestingly, at this fairly low level of proficiency, it is clear that subtest tasks are more closely related to tasks testing the same skill than they are to tasks testing other skills. There is a very clear differentiation between the skills. Most experienced teachers would not find this discovery startling. In the lower and intermediate stages of language acquisition learners clearly develop skills differentially. In other words, a learner may be good at listening and bad at reading. Analyses of tests are different levels of proficiency is reported more fully in Milanovic (1988). The findings of this research indicated that, as learners' language proficiency increased, the skills tended to merge more with each other. A similar finding has been reported by de Jong (1990) using Rasch analysis as opposed to factor analysis. Such evidence casts doubt on the findings of language testing research that does not take the proficiency level of learners into account.

CONCLUSION

The results and procedures described here show that materials-based tests can work. In an educational context, where possible, such tests should be used in preference to approaches further removed from the classroom or real-world context. They are educationally far more desirable than more traditional tests and lose nothing in terms of reliability, if well prepared. In addition, it is time that more innovative tests formed the basis for research in language testing. They would be a more relevant starting point than tests that reflect thinking thirty years ago.

Canale (1985) amongst others, has pointed out that there is often a mismatch between teaching/learning materials and tho : that appear in proficiency -oriented achievement tests. He attributes the mismatch to what he calls the 'image problem', which he breaks down into several categories. First he focuses on the role of the learner in testing and describes him as typically:

> *"an obedient examinee, a disinterested consumer, a powerless patient or even an unwilling victim".*

Canale also focuses on the type of situation that current achievement testing often represents:

> *"... it is frequently a crude, contrived, confusing threatening, and above all intrusive event that replaces what many learners (and teachers) find to be more rewarding and constructive opportunities for learning and use".*

The problems that Canale outlines, which are also of concern to Swain (1985), are major difficulties in the acceptability of testing as an important and useful part of the educational process. Several strategies can be adopted to overcome these problems.

Firstly, testing programmes should be integrated into the life of the institution in which they occur. Testing specialists need to be involved in all stages of curriculum design and not seen as additional extras to the process.

Secondly, the materials used in tests should always reflect the types of activities that go on in the classroom and/or the lives of the students taking the test. In this way both teachers and students will have the better chance of seeing the relevance of tests.

Thirdly, teachers' sometimes inadequate understanding of testing purposes, procedures and principles are often a major barrier in the successful integration of testing into the curriculum in order to overcome this problem, teachers need to be actively encouraged to get involved in test writing projects, and there needs to be a heavy emphasis on their training. Such a strategy not only improves the quality of tests, in te:ms of reliability and validity as illustrated earlier, but also means that more teachers will become familiar with testing as a discipline that is integrated into the education process and not apart from it.

BIBLIOGRAPHY

Alderson, J C., 1983. The cloze procedure and proficiency in English as a foreign language, In Oller, J W (ed), 1983.

Alderson, J C and A Waters, 1983. A course in testing and evaluation for ESP teachers, or 'How bad were my tests?' In A Waters (ed), 1983.

Bachman L F, et al. (1988). Task and ability analysis as a basis for examining content and construct comparability in two EFL proficiency test batteries, Language Testing V5 No 2 pp 128-159.

Canale, M. 1985. Proficiency oriented achievement testing. Paper presented at the Master Lecture series of the American Council on the teaching of Foreign Languages, at the Defence and Language Institute.

Carroll J B. 1961. Fundamental consideration in Testing for English proficiency in foreign students, in Testing the English Proficiency of foreign students, CAL, Washington DC, pp 31-40.

Carroll, B J. 1978. Guidelines for the Development of Communicative Tests, Royal Society of Arts, London.

Carroll, B J. 1980. Testing Communicative Performance, Pergamon Press, Oxford.

Cziko, G. 1984. Some problems with Empirically-based models of communication competance Applied Linguistics, Vol 5 No 1.

Farady, H. 1982. Measures of Language Testing proficiency from the learners' perspective, TESOL Quarterly, Vol 16 No 1.

Hamp-Lyons, Liz (1989a). Applying the partial credit method of Rasch analysis: Language Testing and Accountability, Language Testing Vol 6:1, 109-118.

Milanovic, M. 1988. The Construction and Validation of a Performance-based Battery of English Language Progress Tests Unpublished PhD Thesis, University of London.

Morrow, K. 1979. Communicative language testing: revolution or evolution, In Brumfit, C J and K Johnson (eds), 1979.

Swain, M. 1985. Large-scale Communicative language testing: A case study in Lee, Y.P. et al (eds), New Direction on Language Testing, Pergamon, Oxford.

Wesche, M B. 1987. Second language performance testing: the Ontario test of ESL as an example, Language Testing, Vol. 4, No. 1, 28-47.