NEW DIRECTIONS IN LSP TESTING AT FINNISH UNIVERSITIES

Anna Mauranen
University of Helsinki
Language Centre

Language centres are responsible for the testing of the foreign language proficiency of virtually all Finnish university students. It is important that students are assigned to courses when necessary and that they receive equal treatment in different universities and different times. It is also of particular importance in a specific purpose (LSP) teaching situation that students' needs and opinions are taken into account, and that testing is seen as a positive element which provides useful feedback to students and teachers as well as practical and realistic information to faculties and future employers about students' skills. A testing reform in language centres, started at the Language Centre for Finnish Universities, is now under way, and it attempts to develop the testing principles and ideology at language centres, and eventually to bridge the gap between theory and actual testing practice.

Language centres at Finnish universities have been testing the proficiency of practically the whole student generation in at least two languages for about ten years now. Testing methods and principles have been taken over from mainstream language testing, and these have been given some LSP flavour by using more or less field-specific stimulus materials for the test tasks.

During the course of these ten years the accepted testing ideologies have undergone major changes. In language centres, this has had two consequences: firstly, the standardised tests produced at the Language Centre for Finnish Universities in the early days of language centres have become dated and fallen out of use, and secondly, widely varying practices have emerged among teachers, depending largely on when and where they have received their teaching education, how closely they have followed the developments in the field, or what language centre they happen to be in.

This somewhat anarchistic situation would perhaps be less than ideal in any institution of language education, but in the language centres it undermines the very basis of the system. Language centres are essentially service institutions, which offer teaching required by degree programmes in

university faculties. The role of testing is to find out which students are in need of language instruction in order to meet the degree requirements, and which are not. After language courses, students are tested again, to check that the course has indeed resulted in reaching the set standard. The tests are, then, an operationalization of the language requirement described in degree statutes.

What this system demands from testing is, above all, clear commonly used criteria for required standards and assessment of students. This applies to the initial testing in particular. It is possible to think of alternative goals for final testing, but it is very hard to imagine a situation where the initial tests need not have common goals and common criteria.

Testing reform

In order to develop new solutions to this situation, a testing reform project was started at the Language Centre for Finnish Universities in 1987. The main objectives of the project are, first of all, to achieve improved testing practices, in other words, systematic and meaningful testing, with common principles and criteria. This goal is to be pursued so as to avoid unnecessary work for teachers, and to reduce test-related stress and negative feelings in both students and teachers.

Secondly, the project sets out to gather information about the testing practices in different languages at language centres (see Huhta 1988), as well as to keep up with testing developments in comparable situations in other countries. This includes inviting testing experts from abroad.

The testing reform project also hopes to provoke discussion among language centre teachers about the goals of teaching and learning, and about our implicit or explicit theories of language and what it is to know it and to be able to use a foreign language. This is, in fact, a benefit that will come on the side, since discussions about testing will necessarily touch on questions of how we perceive the nature of language and the objectives of our work.

The testing ideology behind the testing reform builds on the specific demands of Language centre testing (see also Mauranen 1988). Language centres have, at least in principle, adopted an LSP approach to their work. It is this orientation that primarily marks the difference between language centre teaching and general language teaching in schools, or earlier foreign language teaching for university students. In both teaching and testing, this has basically resulted in using relatively field-specific source materials for language tasks. The test tasks themselves, and their assessment, have followed traditional paths in general foreign language proficiency testing. The crucial question for the reform is what other features to testing should be derived from the LSP situation?

The distiguishing feature of LSP teaching is the relatively specifiable goals that students have in their language studies. Because general language teaching usually lacks this knowledge, it naturally aims to test for competence at a very general, abstract level so as to cater for all sorts of possible uses of language. Because in the LSP field we have, in principle, a better grasp of the things our students will want to do with language, we should be in a better position to devise appropriate tests of their ability to carry out these linguistic tasks, and also, hopefully, have better possibilities to gather evidence from actual practice of whether our predictions were justified.

One basis for language centre testing is, then, the students´ language needs. Some needs analyses have been carried out at Language centres in recent years, essentially with the development of teaching in mind (for example, Bullivant et al. 1987, Nordlund 1988). However, needs analyses in their traditional forms of questionnaires and interviews are of limited use for testing, since they are often satisfied with identifying the situations in which languages are used, and the resulting needs descriptions are largely based on the views of the interviewees, in other words people who have no linguistic expertise. Therefore needs analyses tell us very little about what is actually going on linguistically in target situations. In order to find out what the requirements of real situations of language use are, and the problems that speakers encounter, we should observe actual instances of target

communication and analyse typical features of this discourse. Discourse analysis of this kind is relatively close to what Jones (1979) called job analysis, ie. the determination of the skills being tested in relation to the job as a whole. To achieve this, we should rely on both needs analyses and analyses of discourse in target communication situations.

Performance testing

It appears to be a common worry among LSP testers that students' needs are not adequately taken into account. Another major concern is that traditional language tests with their focus on linguistic knowledge are not very good at predicting testees' actual success in real-life tasks (see for example Hauptman et al. 1985). This also frequently seems to apply to tests that are meant to be "communicative", and have been based on notional and functional considerations. In answer to this, some testers and researchers have suggested that , in addition to identifying the needs, the test tasks should be brought as close to the target reality as possible. This trend is often called performance testing, or direct testing.

Performance is not to be understood in terms of the Chomskyan dichotomy of competence and performance. Most people concerned with performance testing are not explicitly trying to measure performance in order to make statements about an assumed underlying competence. The aim is, rather, to try to predict future performance in certain tasks by sampling that performance and assessing the sample. The term performance testing is originally borrowed from the field of vocational testing. One of the proponents of performance testing is Wesche, who defines it in the following way: "[In performance-based tests] examinees must demonstrate their second language proficiency through tasks whose content and contextual features represent the situations in which the second language will eventually be used." (Wesche 1987)

The term performance testing has been used in different ways, but some common features are shared by most interpretations. These are above all authenticity, or realism of all aspects of testing, going beyond narrowly

defined linguistic skills, criterion-referenced testing, interactive test tasks and systematic rater training.

Some performance testers (for example, Emmett, 1985, Bailey, 1985, Weir, 1988) maintain the competence-performance dichotomy and simply believe that direct performance type testing will give a more realistic grasp of a testee's underlying competence than more analytical testing. Others, again, (e.g. Jones 1979, 1985) ignore the concept of competence or are downright hostile towards it (Economou MS) and set out sampling and simulating the kind of performance they will want to predict. This theoretical difference has not been much reflected in the practical solutions adopted but it does affect the kind of statements that can be made in each case about the testees.

According to Slater (1980), there are three basic types of performance testing:

(1) direct assessment, where the examinees are observed in real situations, and thus test tasks are not manipulated at all. Following a student in a foreign country, observing how he or she manages linguistic situations would be an example of direct assessment applied to language testing. Here authenticity is of course perfect, but there is little consistency in the measurement. The practical implementation of measurement of this kind is also cumbersome.

(2) work sample evaluation, which entails some task manipulation, in order to increase comparability, consistency and efficiency of measurement. An instance of this would be assessing students' subject-specific written tasks separately for language.

(3) simulation techniques, which do not involve real situations, but supposedly represent the crucial aspects of real-life contexts. This is the most commonly used and usable alternative in language testing.

One of the major advantages of performance testing is its potentially good backwash effect on teaching: maximal authenticity in tests will encourage teaching which aims at good real-life skills of language use. Performance

tests also give a better chance for students to show their skills than more analytical tests, since they allow for more variation in the skills profiles of individuals. The profiles may be very different even when the achievement of communicative goals is at the same general level. In other words, good communication may be achieved with very different means, and non-linguistic skills may compensate for gaps in linguistic skills. Whereas for teaching purposes it can be argued that it is essential to decompose the behaviour in question, and to practise different aspects of it separately, there is no similar case to be made for testing. When we test, it is important to see whether the testee can put all the different components of the tested behaviour together and apply these to practical tasks. Reading is a good example: we want to know if somebody can arrive at a reasonable interpretation of a text, not whether he/she arrived at it by using the procedures that we had prescribed, be it knowledge of individual grammatical items, or ability to utilise discourse markers, or whatever.

Problems with performance testing

There are also problems involved with performance testing. First of all, even though performance test advocates criticize more traditional tests for lack of predictive ability, the possible better predictive ability of performance tests is yet to be shown. In principle, prediction can be divided into two kinds: time-related and domain-related prediction. The former tries to predict later behaviour with earlier, and can be confirmed with a second measurement. The latter is more problematic: a sample is supposed to predict a whole, for example if our test indicates that somebody speaks Swedish excellently, we assume that this is true of the person's command of Swedish in general. This form of prediction rests essentially on the representativeness of the behaviour sample in relation to the whole behaviour domain that we want to make predictions about, which is very difficult to show empirically. The more limited the domain in question, however, the more feasible it is to assess the representativeness of a sample.

A second problem with performance testing is the impossibility of perfect simulation. This is obvious and in itself need not be a problem, but since

performance testing takes maximum authenticity as its prime principle, it necessarily means falling short of the target. The problem can be solved, however, with a good model of both the communication situation and linguistic behaviour. Simulation is, fundamentally, a kind of model. It is based on an analysis and reproduction of the relevant elements of reality for a given purpose, not the reality as a whole. A model of the communicative situation can be any that adequately defines a situation type, for example in terms of sender, receiver and message, or field, tenor and mode. The kinds of linguistic skills relevant in most communicative situations are relatively well captured in for instance the model of communicative performance suggested by Canale and Swain (1981, and Canale 1984).

The third problem with performance testing is that very careful defining and interpretation of assessment criteria is needed, much more so than in more controlled tests. It follows that it takes a lot of time and effort to train raters and to develop adequate rating scales. What often seems to happen in tests that are meant to be "communicative" is that the student is given a communicative task and then assessed in terms of linguistic knowledge. For example, students can be interviewed, or observed in a discussion, and at the same time evaluated with a rating scale that is largely based on lexicogrammatical or pronunciation errors. In such a test situation, the testee's avoidance of some linguistic expressions, or replacing them by simple roundabout expressions, can be interpreted as indications of gaps in linguistic knowledge, which they undoubtedly in some sense are. In performance terms, however, these can be signs of strategic ability and should thus be evaluated positively. The rating scales and assessment criteria need, then, to be as carefully defined and applied as the test tasks. If they are not, the testing is not performance testing, however well the simulated task reflects target reality.

Fourthly, the most commonly used methods of determining reliability are not applicable to performance tests, since these methods usually assume a large number of independent items, and unidimensionality in the trait measured. However, inter-rater and intra-rater reliabilities are applicable, and it is possible to achieve high reliability values on these measures. Validity rests mainly on construct and content validity, and empirical

validity can also be measured. This differs little from any other kind of testing.

Other features of language centre testing

Besides the basic idea of attempting to derive test content, tasks and criteria from the students´ expected real-life language needs, there are other aspects of testing that need to be considered in the light of the specific testing situation at language centres. An important question is student-centredness. Two of the aspects that Alderson (1986) takes up in connection with student-centredness are relevant in language centre testing. The first is students' own views of how they might be appropriately and 'fairly' tested. This is basically a question of a test's face validity, which, although often belittled in traditional test theories, can nevertheless have a significant role in the test's overall validity through influencing testees' performance level (cf. Low 1985). Authenticity in test situations may in itself contribute to students' accepting the tests as relevant, but students should also increasingly be consulted about their preferred testing methods. The second aspect mentioned by Alderson is students' self-evaluation, which ought to be encouraged in university students. Self-evaluation is used in some language centres, and the present project has also started systematic experimentation in this area at the University of Tampere.

Furthermore, to foster students' positive attitudes towards testing as far as possible, tests should aim at tapping the testees' best performance, and not try to trap them or find their weaknesses. Most teachers will probably agree with this in principle, but in practice it is the common areas of difficulty which often produce most dispersion of test scores and thereby differentiation of students. Consequently these difficult areas will be included in tests, to give well-scattered scores. Since in language centres a dichotomous test result is usually enough, scores need not be scattered all along the scale in the way normally expected with more subtle ranking. There is therefore little need to concentrate on students' problem areas.

It is possible, and indeed desirable, to apply the principles outlined in this paper to a variety of testing methods. Testing situations vary greatly in

terms of size of student population, time and resources allocated for testing, skills areas to be tested, and the general proficiency level of students. It is therefore necessary to maintain a flexible approach to the particular testing methods applied. There are some test formats that are difficult to reconcile with the idea of performance testing, above all multiple choice tests, but a limited application area could be found for even these, in situations where choosing between alternative interpretations of the same input is normally done in real life. Performance testing should not be seen as a fixed task type but rather a set of principles used in preparing tests and analysing them.

Variety in test methods has potentially a good backwash effect on teaching, since it allows for a variety of teaching methods. Moreover, there are two other advantages to combining different methods in the same test or test battery. First, an important benefit is the possibility of different test parts compensating for each other's weaknesses. Different preferences and cognitive or personality styles of students can also be taken into account in a multimethod approach. Secondly, since the sample of student behaviour is drawn from different sources, it gives a wider view of the student's performance potential. A large sample also has the advantage that assessment need not be based on details or individual errors.

In conclusion, it is important to develop testing in language centres towards solutions which are sensitive to the special features of the language centre situation. We should aim at authentic, realistic testing which rests on a careful analysis of students' needs and the requirements of target discourse situations. This approach is made necessary by the fact that we are dealing with LSP; it is also the LSP aspect which makes it possible, by specifying the fields of language use. What still needs to be done, after we know where the languages are to be used, is to find out how the languages are used, to fulfill the specified purposes. This is a vast area which we still know very little about.

References:

Alderson, J.C. 1986. Innovations in Language Testing? In Portal, M. (ed.), Innovations in Language Testing. Windsor: NFER-NELSON. 93-105.

Bailey, K.1985. If I Had Known Then What I Know Now: Performance Testing of Foreign Teaching Assistants. In Hauptman et al. (eds.), 153-180.

Bullivant, D, Lönnfors,P., Nordlund J., Satchell, R. 1987. Needs Analysis for the Lay Person. Language Centre News 9/1987.

Canale, M. 1984. A Communicative Approach to Language Proficiency Assessment in a minority Setting. In Rivera, C. (ed.), Language Proficiency and Academic Achievement. Clevedon: Multilingual Matters.

Canale, M. and Swain, M. 1980. Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing. Applied Linguistics 1/1, 1-47.

Economou, D. 1987. Language Testing in Academic Contexts. Unpublished manuscript.

Emmett, A.1985. The Associated Examining Board's Test in English for Educational Purposes (TEEP). In Hauptman et al. (eds.), 131-152.

Hauptman, P. LeBlanc, R., Wesche, M. (eds.). 1985. Second Language Performance Testing. Ottawa: University of Ottawa Press.

Huhta, A. Outline of English Language Oral Skills Testing in the Language Centres,. Language Centre News 10/1988.

Jones, R. 1979. Performance testing of second language proficiency. In Briere, E.J. and Hinofotis, F.B. (eds.), Concepts in Language Testing: Some Recent Studies. Washington, D.C., TESOL. 50-57.

Jones, R.L.1985. Second Language Performance Testing: An Overview. In Hauptman et al. (eds.).15-25.

Low, G. 1985. Validity and the Problem of Direct Language Proficiency Tests. In Alderson, C. (ed.), Evaluation. Lancaster Practical Papers in English Language Education, Vol. 6. Oxford: Pergamon.

Mauranen,A. Kielitaidon mittaamisesta kielikeskuksissa. Language Centre News, 3/1988.

Nordlund, J. 1988. English Oral Skills for Theatre Academy Students in Finland. MA thesis, University of Birmingham

Slater, S. 1980. Introduction to performance testing. In Spirer, J. (ed.), Performance Testing: Issues Facing Vocational Education, Columbus, Ohio: The National Center for Research in Vocational Education. pp. 3-18

Wesche, M. 1987. Second language performance testing: the Ontario Test of ESL as an example. Language Testing 4/2.