Performance testing for the professions: language proficiency or strategic competence?

Cathie Elder and Annie Brown
The NLLIA Language Testing Research Centre
The University of Melbourne

Abstract

The paper addresses the issue of language performance testing in the professions and focuses particularly on some problems involved in assessing the oral language proficiency of teachers and tour guides.

Three procedures are discussed: a classroom-based observation schedule used to assess the English language proficiency of nonnative secondary school teachers of maths and science, a less direct test of Italian language proficiency with tasks designed to simulate the demands made of foreign language teachers in the classroom situation, and an oral test of Japanese for tour guides consisting of a series of simulated occupational tasks.

A key question arises in relation to each of these assessment procedures: is it feasible or valid to assess language proficiency independently of the strategic/pragmatic behaviours required for successful performance in the occupational context? Evidence from the trialling of the above procedures suggests that when raters with relevant occupational expertise are involved in the assessment process there may be a conflict between the assessment of language proficiency as traditionally conceived and the evaluation of communicative competence in relation to the particular requirements of the professional situation. Samples of test discourse, qualtitative feedback from raters and analysis of rating patterns of raters from different backgrounds are used to illustrate this point, and implications are drawn for performance testing of language skills in other occupational areas.

Introduction

Performance-based language testing is now a widely-accepted form of testing, popular perhaps because of its obvious relevance and utility, particularly where assessment is being carried out in relation to specific occupational contexts. But the attempt to bring the real world into the testing situation carries with it problems of construct definition: to what extent are we dealing with language rather than other factors when making judgements about individual candidates?

Melbourne Papers in Language Testing 1997 Volume 6.1 pp. 68-78. The Language Testing Research Centre, The University of Melbourne.

The influence of factors other than language in performance-based language assessment has long been acknowledged by language testers (see, for example Jones, 1985; McNamara, 1990; Wesche, 1992; McNamara 1996). While some writers take the Hymesian view that these non-linguistic factors, such as sensitivity to audience, interactive skill and personal style, are part and parcel of communicative competence, others see them as beyond the scope of language testing or a source of what Messick (1992) describes as 'construct-irrelevant variance'. In discussing this issue McNamara (1990) makes an interesting distinction between 'strong' performance tests, in which test tasks are the *target* of the assessment with language being treated as a necessary but insufficient condition for their successful execution, and 'weak' performance tests, in which language proficiency is assessed independently of other factors involved in the performance, and tasks serve merely as *vehicles* for eliciting a relevant language sample.

Nevertheless, any attempt to simulate demands of particular occupational contexts will invite (explicitly or implicitly) consideration of aspects of strategic competence, in Bachman's terms 'the capacity that relates language competence, or knowledge of language, to the language user's knowledge structures and the features of the context in which communication takes place' (1990: 107) and which 'enables an individual to make the most effective use of available abilities in carrying out a given task' (1990: 106). It is doubtful therefore whether in practice McNamara's weak/strong distinction can be maintained, since what ends up being assessed may depend less on principled decisions made at the test design stage than on the way candidates manage the particular requirements of the testing situation and upon the particular orientation of the raters involved in the assessment process.

In this paper we discuss the role of strategic competence in relation to three occupation-specific language tests. All three tests are situated towards the "strong" end of the performance test continuum in that they explicitly invite judgements about the effectiveness of task performance in relation to the demands of the real world context as well as about the quality of the language sample. The inclusion of aspects of strategic competence in the assessment criteria was considered important in all three cases because our test users needed to know whether those being certified could adjust to changing situational conditions, and tailor their language performance to the context.

The Japanese Tour Guide Test (see also Brown 1994, 1995) and the oral component of the Italian Teacher Test are similar in format in that in both candidates are required to complete a series of simulated occupational tasks taking on the relevant professional role (tour guide or language teacher). Candidates are assessed on two broad aspects of the performance, linguistic skill and task fulfilment or communicative competence.

The third procedure is designed to monitor the English proficiency of graduates from non-English-medium universities who are training to be teachers of maths and science (see also Elder 1993, 1994). Unlike the previous two tests it doesn't involve simulation but is an observation schedule, administered in the classroom by maths/science teachers and teacher-trainers while the trainee is conducting a practice lesson. Performance indicators are grouped under six headings the first five of which can be regarded as components of language proficiency, with the last being specific to the classroom context and falling within the parameters of Bachman's 1990 definition of strategic competence. An overall assessment of communicative efectiveness is also made.

The remainder of this paper will focus on two problems which arose out of the attempt to assess strategic competence in the test situation: that of defining the trait and that of choosing suitable raters to undertake the assessment.

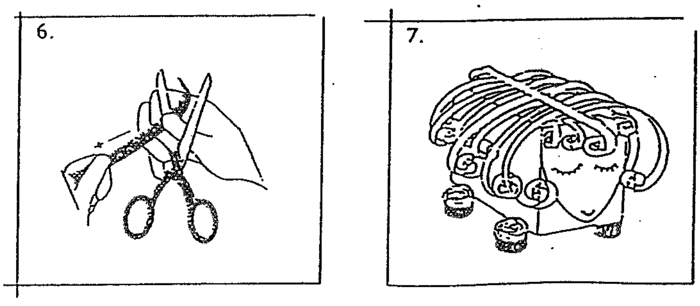
Problem A: Defining and measuring the trait

The problematic nature of the *task fulfilment* criterion on the Tour Guide Test emerged during a workshop convened with tourism industry and Japanese language teacher informants. Whilst there appeared to be general agreement as to the level of performance of particular candidates on particular tasks, it was not possible to specify for all candidates and for all tasks what aspects of performance should be considered in reaching the judgement. In other words, the trait addressed by the *task fulfilment* criterion seems to be less clearly definable, being both multi-faceted and variable, and drawing on a range of performance features which combine and compensate in ways which are neither fixed nor predictable in advance. This of course presents a problem in so far as providing meaningful test reports and explicit direction to raters is concerned.

In the analysis of rating patterns on both tests it appeared that, as was to be expected, there was some correlation between assessments on the 'linguistic' criteria and those on the 'communicative' skills. However, the relationship was not such that they could be considered to be measuring the same trait. In the Teacher Test in particular, there were at times sizeable discrepancies between ratings on the two types of criteria for individual candidates.

It was found that there were large numbers of misfitting ability estimates in the output which indicated that 12% of the 75 candidates appeared to have these unexpected patterns of ability, ie high scores on the linguistic criteria amd low on the communicative criteria, or vice-versa. In an attempt to find out the possible source of these 'disorderly' measurements, transcriptions of test discourse were undertaken for a sample of the misfitting candidates. Presented below is a short segment of performance from a representative of each of these two groups of candidates. These are taken from performance on an instruction-giving task in which candidates are given a set of picture prompts and are asked to explain, *as they might to a group of young second language learners,* how to make a paper model of a sheep.

Figure 1: Instruction-giving task



CANDIDATE A (HIGH on classroom competence LOW on linguistic competence)

> **dovete fare...così e.. ecco ...avete il piede della pecora**
>
> you have to do ...like this... and...here... you have the sheep's foot

CANDIDATE B (LOW on classroom competence, HIGH on linguistic competence)

> **per formare le zampe del pecorello si prendono le striscie di carta e con le forbici**
>
> to make the sheep's hooves you take the strips of paper and with the scissors
>
> **le si ... arrotolano**
>
> you... roll them up

In describing one step of the activity, Candidate A demonstrates what has to be done by showing with her hands the action of curling a strip of paper with a pair of scissors, accompanying this with a somewhat limited linguistic description. Candidate B, on the other hand, describes the action with words rather than gestures, using more sophisticated syntax and more precise lexis.

It is easy to see why candidate B (who is in fact a native speaker of Italian) was awarded a high rating for *linguistic* competence but a relatively low mark for *classroom* competence and why the opposite was true for the first candidate. (We should point out here that candidate A's low score on linguistic competence is a result of performance across all tasks). Although candidate A's utterances *may* be a direct result of her lack of linguistic competence, she has used language which is arguably more appropriate for young second learners with limited linguistic proficiency.

This example draws attention to what may be a fundamental incompatibility between the traditional notion of general proficiency which assumes a developmental continuum involving an increase in range and complexity of language use across all contexts, and the nature of performance in specific situations where attributes such as simplicity, clarity and sensitivity to audience may be valued over and above elaborateness. On a test such as this one it seems that on certain tasks we have a clash of 'frames' and that linguistically proficient learners, understandably anxious to 'show off' their level of linguistic sophistication, are sometimes outperformed on certain dimensions of assessment by less proficient speakers who respond (whether consciously or unconsciously) more appropriately.

Given that both these tests are used for selection purposes, there are, furthermore, issues of social equity to be considered. Is it reasonable to measure, and hence demand, skills of one group of people (nonnative speakers in the case of the tour guide test) and non-graduates in the case of the teacher test) that others are not required to demonstrate? By measuring a trait which could be expected to develop as a a result of job experience, are we, moreover, disadvantaging those candidates who may have an adequate linguistic basis but are professionally relatively inexperienced? Furthermore, given the difficulty of rèplicating the contextual features of the Italian language classroom or the Australian tour in a test environment, can we regard an unconvincing role simulation on our test as indicative of inability to perform in the real world?

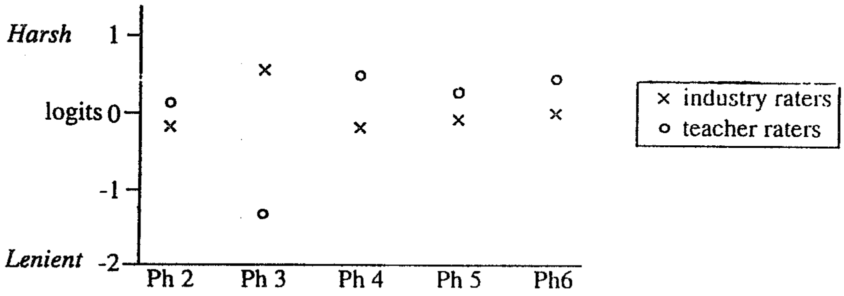
For both tests our practical solution to these concerns has been to separate classroom competence ratings from linguistic ones in reporting performance. For the teacher test it is advocated that the classroom competence ratings be used only for diagnostic purposes and not for selection except in borderline cases, and for the tour guide test the final grade awarded reflects the linguistic criteria only, with a subsidiary statement in relation to the task fulfilment criterion. This ensures that grades are not unduly influenced by any failure to fulfil the contextual demands of the task for whatever reason, be it limited experience of the professional situation or a lack of ease at being required to 'act' in a test situation, but, on the other hand, that information is also available about those candidates who may be linguistically weak, but are able to compensate through good strategic or communicative ability.

Problem B: Choosing the Rater

We explored the orientations of raters from these different backgrounds to the assessment criteria by using data drawn from the rater accreditation process of the Tour Guide Test, where 13 raters had a tourism industry background and 9 a Japanese teaching background.

In relation to the 'task fulfilment' criterion, the industry raters were in general marginally more lenient than the teacher raters with the exception of Phase 3 (dealing with an upset or worried client). On this phase teachers were more lenient than they were on any of the other phases, whereas industry raters were much harsher.

Figure 2: The Japanese Test for Tour Guides: Rater differences



During the test development stage, industry informants had expressed the view that the ability to cope with such a situation (dealing with an upset/worried/angry client) was a crucial skill in guiding. It seems, then, that the importance they place on this task is reflected in industry raters' assessments: they score candidates harshly on this phase, an indication that they are not prepared to tolerate inadequate performance. Teachers, on the other hand, are much more generous on this phase, possibly because they perceive highly-charged exchanges which require great diplomacy as being particularly difficult from the point of view of learning the language

Similar variability emerged in relation to the The Classroom Language Assessment Procedure. A study comparing judgements made by 8 subject specialist raters to those of 7 ESL teachers, showed considerable differences in orientation. There was an unacceptably low level of agreement between the two groups in their assessment of 'subject specific language use'. Feedback suggested that this was because the ESL teachers were focusing on the lexis, grammar and internal cohesion of the candidate's presentation while the subject specialists were more concerned with the way in which subject content was conceptualised.

A stepwise regression was carried out on the data provided by the two groups of raters in order to explore which aspects of language use contributed the most to the global assessment category 'overall communicative effectiveness'. For the ESL raters, 'subject specific language use' emerged as the first variable and 'comprehension' (that is, ability to comprehend the learners) as the second. In contrast, the subject specialist data selected 'classroom interaction' only. This category is concerned solely with features of strategic competence, which sit less comfortably with the commonly held view of what constitutes proficiency. It seems therefore that ESL raters pay more attention to aspects of linguistic skill than subject specialists, who are more concerned with classroom behaviours. That the subject specialists are somewhat ill at ease with language matters was also borne out by the relatively low levels of intragroup reliability achieved by the subject specialists on the linguistic criteria.

Table 1: Maths / Science Observation Schedule

Relationship between analytical and 'overall communicative effectiveness' scores

Step	Category	R^2	Change in R^2	t
ESL raters				
1.	Subject specific	69.53	69.53	5.45**
2.	Comprehension	82.54	13.01	2.99**
Subject specialist raters				
1.	Interaction	94.57	94.57	16.69**

**p< 0.01

Discussion

The issues discussed here raise a number of important validity questions in relation to performance testing for the professions.

- *How should the various traits underlying language use in real world contexts be defined?*

As we have seen, teachers and occupational experts appear to operate from different schemata in judging test performance. While it is generally accepted that occupational experts should be consulted at the needs analysis phase of test development in regard to task design, assessment criteria are generally stipulated and weighted by the test developer, a linguist, with reference (at best) to theoretical models of language ability or to commonly accepted and used assessment frameworks rather than with reference to real-world judgements. We pose the question here of the extent to which such an approach to test development, one which marginalises the contribution and views of industry members, provides valid and useful information.

- *Who is the best judge of job-related performance?*

It could be argued that occupational experts are linguistically naive and less reliable and that it is therefore inappropriate to entrust them with the task of assessing language-related behaviours. The findings reported above however suggest that language teachers may not be the most appropriate judges because they are less inclined or less able to focus on relevant *context-specific* skills. Indeed, if we accept that there are instances of language performance where the formulation of an acceptable and intelligible message depends on occupation-specific knowledge or expertise, then the involvement of occupational-experts as assessors should be regarded as a condition of test validity.

- *How should test performance be reported?*

Once the traits to be assessed have been defined and suitable judges have been chosen, we are left with the question of how the traits should be reported in relation to one another. For the tests considered here we have devised some practical solutions in which information about candidates' performance against task fulfilment or classroom competence criteria is treated as subsidiary diagnostic information, is used to assist in decisions about borderline cases, or provides data on skills which may compensate for linguistic shortcomings.

In theoretical terms, however, this solution amounts to a weakening of the test's claim to specificity. If information about general language proficiency is enough, there seems to be little point, other than satisfying the need for face validity, in trying to capture the context-specific features of language performance. If, on the other hand it can be demonstrated that in a particular context these strategic skills are as important as language proficiency, then there may be a case for giving them greater weight in the assessment process. But we need first to be sure that criteria used to measure such aspects of performance are indeed measuring relevant skills (all the more so if, as was the case with the Italian test, these skills appear to be at odds with language proficiency as traditionally conceived). This may be easier to ascertain with 'on the job' assessments such as the classroom-based assessment procedure, but with less direct measures, where the candidate is required to simulate the professional role, there is, as we have suggested, a risk of measuring construct irrelevant facets of the test method (such as acting ability, or ability to 'suspend disbelief' in a roleplay situation). Rather than being seduced by the appearance of authenticity into accepting that performance tests are necessarily more valid than traditional types of assessment, we need to find ways of ensuring that there is a reasonable degree of fit between behaviours elicited from candidates in the artificial environment of the test and actual performance in the target domain.

References

Bachman, L.F. (1990) Fundamental considerations in language testing. Oxford: Oxford University Press.

Brown, A. (1994) LSP Testing: The role of linguistic and real-world criteria. In Khoo, R. (Ed.) LSP: Problems and Prospects. Singapore: SEAMEO-Regional Language Centre.

Brown, A. (1995) The effect of rater variables in the development of an occupation-specific language performance test. Language Testing 12,1: 1-15.

Elder, C. (1993) How do subject specialists construe classroom language proficiency? Language Testing 10, 3: 235-254.

Elder, C. (1994) Are raters' judgements of teacher proficiency wholly language based? Melbourne Papers in Language Testing 3,2: 41-61.

Jones, R. (1985) Second language performance testing: an overview. In Hauptman, P., Le Blanc, R. and Wesche, M. (Eds.) Second language performance testing. Ottawa: University of Ottawa Press: 15-24.

Messick, S. (1992) The interplay of evidence and consequences in the validation of performance assessments. Princeton, NJ: Educational Testing Service.

McNamara, T.F. (1990) Assessing the second language proficiency of health professionals. Unpublished PhD, The University of Melbourne.

McNamara, T.F. (1996) Second Language Performance Assessment. London: Longman.

Wesche, M. (1992) Performance testing for work-related second language assessment. In Shohamy, E. and R. Walton (Eds.) Language assessment for feedback: testing and other strategies. Washington, D.C.: National Foreign Language Center Publications.