Assessing the language proficiency of teachers: are there any border controls?

Catherine Elder

Language Testing 2001; 18; 149
DOI: 10.1177/026553220101800203

Assessing the language proficiency of teachers: are there any border controls?

Catherine Elder *University of Auckland*

This article takes up some of the issues identified by Douglas (2000) as problematic for Language for Specific Purposes (LSP) testing, making reference to a number of performance-based instruments designed to assess the language proficiency of teachers or intending teachers. The instruments referred to include proficiency tests for teachers of Italian as a foreign language in Australia (Elder, 1994) and for trainee teachers using a foreign language (in this case English) as medium for teaching school subjects such as mathematics and science in Australian secondary schools (Elder, 1993b; Viete, 1998).

The first problem addressed in the article has to do with specificity: how does one define the domain of teacher proficiency and is it distinguishable from other areas of professional competence or, indeed, from what is often referred to as 'general' language proficiency? The second problem has to do with the vexed issue of authenticity: what constitutes appropriate task design on a teacher-specific instrument and to what extent can 'teacher-like' language be elicited from candidates in the very artificial environment of a test? The third issue pertains to the role of nonlanguage factors (such as strategic competence or teaching skills) which may affect a candidate's response to any appropriately contextualized test-task and whether these factors can or should be assessed independently of the purely linguistic qualities of the test performance.

All of these problems are about blurred boundaries, between and within real world domains of language use, between the test and the nontest situation, and between the components of ability or knowledge measured by the test. It is argued that these blurred boundaries are an indication of the indeterminacy of LSP, as currently conceptualized, as an approach to test development.

I Background

Douglas (2000: 46) defines a specific purpose language test as:

> [A test] in which the test content and methods are derived from an analysis of the characteristics of the specific target language use situation, so that test tasks and content are authentically representative of the target situation, allowing for an interaction between the test-taker's language ability and

Address for correspondence: Department of Applied Language Studies and Linguistics, University of Auckland, Auckland, New Zealand; e-mail: c.elder@auckland.ac.nz

Language Testing 2001 18 (2) 149–170 0265-5322(01)LT199OA © 2001 Arnold

> specific purpose content knowledge, on the one hand, and the test task, on the other.

In theory at least, such tests allow us to make more accurate inferences about a test-taker's future performance in the relevant target language use (TLU) domain than is the case for a test of general language proficiency where the criterion domain of reference is an abstraction rather than a reality.

Douglas (in press), however, acknowledges a number of problems which dog the enterprise of testing Language for Specific Purposes (LSP). The first has to do with specificity: how one defines the domain of reference and distinguishes it from other real world domains. The second has to do with authenticity: how one ensures that a test reflects the demands of the real world situation and elicits an appropriate sample of language from the test-taker. The third pertains to the role of nonlanguage factors (which inevitably influence performance on any appropriately contextualized test-task) and whether these factors can or should be assessed separately from the purely linguistic aspects of test-taker behaviour. The three problems identified by Douglas all have to do with boundaries; they are about the boundaries:

- between and within real world domains of language;
- between the test and the nontest situation; and
- between the components of ability or knowledge measured by the test.

This article explores these three problem areas of LSP testing and the challenges they pose for assessing the language proficiency of teachers. The problems will be illustrated with reference to three specific purpose test instruments:

1) A test of English for the selection of overseas-qualified immigrants applying for entry to teacher education programs in the Australian state of Victoria. This test is known as the Diploma of Education Oral Interview Test of English (Viete, 1998) (hereafter referred to as DOITE).
2) A classroom language assessment schedule designed to identify the English language problems faced by nonnative speaker teachers in training during their school-based teaching practice rounds (Elder, 1993b) (hereafter CLAsS).
3) Language proficiency test for teachers of Italian and Japanese as a foreign language in Australia (Elder, 1994; Elder *et al*., 1995) (hereafter LPTT)

1 DOITE and CLAsS tests

Both DOITE and CLAsS are intended for overseas-trained subject teachers in specialist areas, such as mathematics or science, without a locally recognized qualification in teaching methodology (Viete, 1998). The DOITE is used for selection purposes, to determine whether the English language skills of overseas applicants for secondary teacher education are sufficient to allow them to cope with the demands of their training course. The CLAsS follows on from the DOITE and is designed for use on the school-based teaching-practice round during which teacher trainees are required to teach their subject specialism to secondary students under the supervision of a registered teacher. The purpose of CLAsS is to

1) assist supervisors of successful applicants, who are generally subject specialists rather than language experts, in determining whether any problems they experienced in the teaching situation are related to language; and
2) provide information about aspects of their performance which could be improved through supplementary English language support.

2 LPTT

The LPTT tests, on the other hand, are for language teachers. They were designed for the purposes of professional accreditation, to ensure a minimum standard of language ability amongst graduates with a foreign language major (or equivalent qualification) in Italian or Japanese seeking registration as foreign language teachers. The decision to be made about test-takers is whether their skills in the target language are sufficient to allow them to teach this language to prospective students and to use it as the medium of classroom instruction where appropriate.

3 Type of test

For all the above tests a performance-based LSP model of testing was chosen as the best option. In the case of CLAsS, which is an observation schedule rather than a test instrument proper, the performance is a particular teaching event which takes place in the classroom under the supervision of a qualified teacher. In all other cases the tests attempt to elicit key features of teacher language proficiency via tasks which simulate various aspects of the target teaching situation.

A specific purpose assessment model was chosen for the following reasons:

1) In the case of the foreign language teachers there were concerns (see, for example, Nicholas, 1993) that the language training which the teachers had received in the context of their undergraduate studies was too general or too academic and did not equip them with either the discourse or pragmatic competence necessary to cope with classroom communication.
2) Measures currently used to assess language proficiency of graduates aspiring to teach foreign languages or other school subjects through the medium of a foreign language were considered inadequate predictors of performance in the professional domain. Of particular relevance to DOITE and CLAsS was a study conducted by Elder (1993a) which revealed that the International English Language Testing System (IELTS) scores of overseas students entering teacher-training institutions correlated poorly with their performance on their teacher training course. This was particularly true of the speaking component, which, according to Viete (1998) lacked construct validity for teaching and teacher education purposes.

II Characterizing teacher language proficiency: the problem of specificity

What, then, were the specific skills required for teaching purposes? Needs analysis conducted by Viete (1998) and Elder (1994) revealed that teacher language proficiency was far from being a well-defined domain relying on highly routinized language and a generally accepted phraseology such as is the case with, for example, the language of air traffic controllers (Teasdale, 1994). Instead, it was found to encompass everything that ‘normal’ language users might be expected to be able to do in the context of both formal and informal communication as well as a range of specialist skills. These specialist language skills include command of subject specific/metalinguistic terminology, on the one hand, and the discourse competence required for effective classroom delivery of subject content, on the other hand. Effective classroom delivery necessitates command of linguistic features. Directives are one such feature and are crucial in establishing classroom procedures and learning tasks. A range of questioning techniques is also essential if the teacher is to be able to monitor learner understanding. The teacher will also need to use rhetorical signalling devices and simplification strategies to communicate specialist areas of knowledge and render them comprehensible to learners.

As far as these specialist skills are concerned there is a large body of literature on the characteristics of teacher talk which can assist with domain definition as well as research on the discourse of different subject areas (such as mathematics and chemistry) and their different manifestations in textbooks and lectures. However, as Stevenson (1985: 43) pointed out, we still have insufficient information about the categorical or elemental interrelationships and weights within and among these particular genres to be able to model them accurately and sample from them in a systematic fashion. Moreover, when different genres are combined in different configurations within a larger domain like teaching or classroom communication, this problem is compounded.

Faced with problems such as these in LSP testing we have no other option than to settle for an expedient solution to domain definition (rather than a theoretically defensible one) and to compromise the absolute validity of an individualized test which mirrors precisely a particular target language event but has no generalizability beyond that event, in favour of some degree of abstraction (while at the same time retaining, albeit in simplified or idealized form, something of the communicative 'flavour' of the relevant TLU domain). This trade-off between the real and the ideal is part of the design challenge which all LSP tests must rationally address. Table 1 shows us where this compromise has led us as far as the skills/language functions targeted in the listening/speaking components of the tests nominated above are concerned.

These common elements suggest that, in spite of what seemed an impossibly broad TLU domain, and in spite of differences in what is taught, where and how, there is a high level of consensus about what language functions are critical for teachers across languages and subject areas. Moreover, a similar gamut of skills features in the specifications of teacher tests developed in other countries and based on

Table 1 Language functions assessed in the three tests of teacher proficiency

Skills	DOITE	CLAsS	LPTT
Present information and explain subject specific metalinguistic concepts	✓	✓	✓
Extract meaning from multi-way discussion (with two or more speakers)	✓	✓	✓
Discuss a problem/express opinion	✓	✓	✗
Summarize/paraphrase simplify/disambiguate information	✓	✓	✓
Formulate questions	✓	✓	✓
Issue directives, set up a classroom activity	✓	✓	✓

independent needs analyses (e.g. Hinofotis *et al.*, 1981; Plakans and Abraham, 1990; Grant, 1997).

But do these tests, for all their ‘teacherliness’, really elicit a sample of language that is different in essence from what would be elicited from a general proficiency test? Within-test comparisons of the performance of specialist vs. nonspecialist test-takers can be revealing. Elder *et al*. (1995), for example, when trialling the LPTT Japanese (referred to above), found that expert teachers of Japanese as a foreign language (JFL) (who should have been equipped with precisely the skills the test was designed to measure) performed more poorly overall than did recent graduates from generalist Japanese language courses. While this outcome was doubtless due to the superior language skills of the recent graduates, it nevertheless casts some doubt on the validity of the LPTT for its intended specific purpose.

Between test comparisons (between general and specific measures) tend also to produce equivocal findings. Smith (1992), comparing the SPEAK test designed to test the classroom competence of International Teaching Assistants in the USA with a subject-specific version of it known as MATHSPEAK, found that those with specialist training in mathematics did not necessarily perform better on the version which had been tailored to their needs. Douglas and Selinker (1993) attribute Smith’s findings to the lack of contextualization cues required to promote differential ‘domain engagement’ (p. 235). Contextualization cues mentioned by Douglas and Selinker were changes in tempo, pitch, stress, intonation, volume, overt rhetorical markers, gesture and eye contact, ‘which signal shifts in the dynamics of the communicative event, promoting the planning, execution and assessment of communicative moves and consequently the marshalling of interlanguage resources brought to bear’ (p. 236). The tasks on this test are, in other words, not specific enough to elicit the language behaviour characteristic of the particular criterion domain. But how specific should we be? In domains like teacher proficiency where the range of ‘allowable contributions’ (Swales, 1990: 52) is probably infinite, we have no principled basis for deciding which of the many features of the target context we must sample to be sure that ‘test tasks and content are authentically representative of the target situation’ (Douglas, 2000: 46).

III Simulating the teacher role: the problem of authenticity

Bachman (1991) has made a useful distinction between two different aspects of authenticity:

1) situational authenticity, which refers to the level of correspondence between the test and the TLU situation; and

2) interactional authenticity, which refers to the capacity of the test-task to engage the relevant language abilities of the test-taker.

Situational authenticity (later referred to by Bachman and Palmer as just plain 'authenticity') has to do with task design and is usually established *a priori* by the test-developer. Interactional authenticity (elsewhere referred to as 'domain engagement' by Douglas and Selinker (1993) or 'interactiveness' by Bachman and Palmer (1996)) can only be established after the test has been administered or trialled and candidates' test performance has been analysed.

Situational authenticity has to do with task design and is usually established *a priori* by the test-developer. While we may speculate about the interactional authenticity of a test, or its capacity to engage the test-taker, this can only be established after the test has been administered or trialled and after test performance has been analysed.

I now briefly evaluate the tests under consideration in terms of these two aspects of test authenticity, with particular reference to the way the candidate's role is construed on the test and how this relates to the role of the teacher in the corresponding real world situation. In doing this I am referring to a number of items on Bachman and Palmer's checklist of TLU characteristics which they propose as a basis for matching the test-task to the real world domain of reference, namely: the participants, the channel of the input and expected response and the nature of the input response relationship.

1 CLAsS

CLAsS (see Appendix 1), because it is an observation schedule applied to particular instances of classroom performance, is difficult to fault in terms of situational authenticity. The test-task in this case is a lesson. The trainee teacher is acting out the role of a real teacher in as natural a context as is possible with a real class of learners in front of her. The observer sits at the back of the class and rates different dimensions of the trainee teacher's performance (i.e., fluency, accuracy, comprehension, use of subject-specific language, classroom interaction) but does not intervene in any way during the lesson. In other words there is a perfect correspondence between the assessment situation and the TLU situation (although it is, of course, conceivable that the candidates may modify their normal teacherly behaviour because they are conscious of being observed and assessed).

The interactional authenticity of the instrument depends, however, on what tasks the trainee teacher under observation happens to be performing. Since some classes involve more teacher talk than others, it may be necessary to apply the schedule repeatedly in order to elicit an adequate sample of the trainee's language in a variety of teaching modes. The test's interactional authenticity also depends crucially on the relevance of the criteria used for assessment (i.e., they must be clearly related to the construct of ability which the test is designed to elicit) and on the capacity of the subject specialist teacher charged with making judgements about trainees' language performance to apply these criteria appropriately. (The issue of assessment criteria and the way in which these are applied is revisited below.)

2 LPTT

The LPTT (Italian), while situationally less authentic than the CLAsS because it is not administered in a school classroom, is nevertheless a 'live' communicative performance. As can be seen from the brief description of the test offered in Appendix 2, the candidate is required to assume the role of a primary school language teacher from Phase 2 of the interview until the end. During the course of the test he or she is required to perform various classroom-like tasks, such as reading a story aloud as if to a group of young school-age second language learners, and issuing a set of classroom-like instructions explaining how to carry out a particular learning activity. The quality of candidates' performance on these tasks is assessed using both linguistic and task fulfilment criteria. The task fulfilment criteria draw the assessors' attention to features of communicative behaviour, such as style of delivery, which are deemed to be of particular relevance to classroom performance.

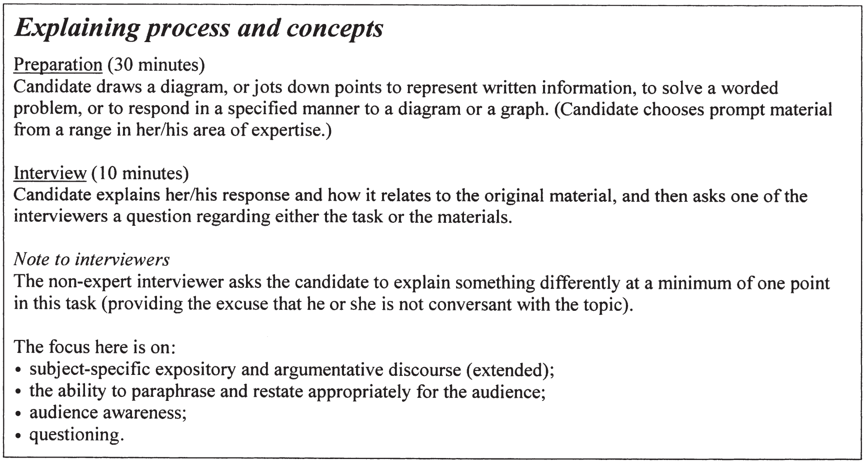
There are, nevertheless, obvious limitations to the test's situational authenticity, the most obvious being that many of the tasks in the oral interview are delivered as a monologue by the candidates. The decision to limit the interactiveness of test-tasks was due to the fact that the candidates could not be expected to communicate with the native or near-native speaker examiners in the same way as they would with a classroom of second language learners with limited control of the target language (Elder, 1993c). It therefore seemed more appropriate for candidates to address an imagined classroom audience and to try to forget the interviewer's presence. The monologic tasks have the added advantage of allowing the candidates, rather than the interviewers, to take control of the talk (as teachers tend to do) and to produce extended stretches of discourse.

Interestingly, the feedback gathered from test trials indicated that it was the monologic tasks, rather than the interactive role plays, which were preferred by both candidates and assessors as measures of teacher proficiency (Elder, 1994). This suggests that it is easier to sustain the illusion of the candidate as teacher when there is minimal input from the interviewer. Whether the sample of language elicited could be regarded as interactionally authentic is of course another matter.

3 DOITE

Because the DOITE test is used to select amongst applicants for teacher education, it is designed to assess candidates' ability to cope with linguistic demands of the Diploma of Education course as well as their ability to cope with their future role as teachers in English-medium classrooms. Test-tasks are therefore designed to fit these two bills simultaneously. However, this dual function of the test places limits on the test's situational authenticity. Presentations on the concepts and processes of mathematics might, for example, be required in both the teacher education tutorial and in the secondary school classroom, but the audience for the presentation would in each case be very different, as would the power relationship between presenter and listener. The following task description (see Figure 1) is fraught with ambiguity.

Although it is fairly clear that this task has been set up to test the candidate's ability to talk about his or her subject specialism in a classroom-like or tutorial-like context, the status of and function of the interviewer input is unclear. One interviewer (possibly in the capacity of supervisor) provides an answer to the candidate's question



Explaining process and concepts

Preparation (30 minutes)
Candidate draws a diagram, or jots down points to represent written information, to solve a worded problem, or to respond in a specified manner to a diagram or a graph. (Candidate chooses prompt material from a range in her/his area of expertise.)

Interview (10 minutes)
Candidate explains her/his response and how it relates to the original material, and then asks one of the interviewers a question regarding either the task or the materials.

Note to interviewers
The non-expert interviewer asks the candidate to explain something differently at a minimum of one point in this task (providing the excuse that he or she is not conversant with the topic).

The focus here is on:
- subject-specific expository and argumentative discourse (extended);
- the ability to paraphrase and restate appropriately for the audience;
- audience awareness;
- questioning.

Figure 1 Sample task description from DOITE (Viete, 1998: 184)

about a problem and the other (possibly in the role of student) asks the candidate for an explanation or reformulation of his or her input at one or more points in the presentation.

However, even if the role relationships were made more explicit, and the relationships between the task characteristics and the criterion domain of reference were more clearly delineated, there is no guarantee that the resultant performance by the candidate would be interactionally authentic. Interactional authenticity would be dependent on both candidate and interviewer conforming to the requirements of their respective roles (for example, the nonexpert interviewer's request for a reformulation would need to be made in a 'natural' and nonintimidatory manner). Lumley and Brown (1996), in their analysis of nurse–patient role play in a test of medical English, have pointed to the possibility that interruptions from the interlocutor – even when these are perfectly in keeping with the patient role that they are assuming – may inhibit the candidate to a point where he or she is unable or unwilling to continue the interaction and may therefore fail to produce an assessable sample of speech.

In sum, while the three tests cover similar ground in terms of the skills they purport to measure, the framing of test-tasks and the environment in which they are administered varies considerably from test to test. And while some tests appear more situationally authentic than others, it is not clear how important this is for the interactional authenticity of the test. It goes without saying that any simulation of the teacher role in the test situation requires considerable suspension of disbelief on the part of all the participants, i.e., the candidate, the interlocutor and the rater. Attempts to give authority to the candidate in the role of teacher or knower or to wrest it away from the interviewer may be counterproductive, since it will be quite clear to all concerned that the 'teacher' has no real authority to draw on nor a classroom reality to refer to (except of course in the case of the ClasS where the performance is, in fact, a teaching event). Likewise an interviewer who assumes the role of student or naive listener, judge and facilitator of communication will simultaneously be unconvincing, at least for some test-takers, and the more sceptical among them may end up being penalized for their inability or unwillingness to conform to the script.

The general point to be made is that, as Bachman and Palmer (1996: 174–75) and Douglas (2000: 128) also point out, the requirement of test usefulness and practicality often necessitates modification of task characteristics and conditions at the expense of situational authenticity. The effect of such modifications on the quality of language performance and on the interpretations we place on the meaning of these performances in relation to the real world domain of reference is far from being fully understood. Further research is needed (1) to map the relationship between the discourse produced in real world domains and the language behaviour elicited on test-tasks expressly designed to mirror these domains and, where possible, (2) to determine which features of the testing context are producing observable disparities between test and real-world discourse. Ideally, this kind of research should be part of the normal iterative process of test validation such that the design blueprints for LSP tests are seen as dynamic rather than static and features of the testing situation can be modified on an ongoing basis to bring them more closely into line with the characteristics of the relevant TLU domain.

IV Language ability or classroom competence?: the problem of inseparability

To the extent that we succeed in producing tests which are sufficiently context sensitive to elicit a valid sample of 'domain specific' behaviour from the candidate, we face a third boundary problem: that of inseparability. Should we, and indeed can we, assess the contextual features engaged in language performance independently of the language sample itself?

The influence of factors other than language in performance assessment has long been acknowledged by language testers (e.g. Jones, 1979; Wesche, 1992; McNamara, 1996; Jacoby and McNamara, 1999), and while some writers take the view that these nonlinguistic factors such as sensitivity to audience and personal style are part and parcel of communicative competence, others see them as being beyond the scope of language testing or a source of what Messick (1993) describes as construct-irrelevant variance. In discussing this issue McNamara (1996) makes a 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 of 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.

He advises against the strong approach, which focuses on the single test performance, because this limits the generalizability of test-scores. The weak approach – because it is more concerned with underlying language skills engaged in performance than with the qualities of the performance in its own right – allows us to make inferences about a wider range of future performances in the real world.

In practice, however, it is doubtful whether McNamara's weak/strong distinction can be maintained in a performance-based LSP test since what ends up being assessed may depend on the way candidates and their interlocutors manage the particular requirements of the testing situation and/or upon the particular orientation of raters involved in the assessment process. On the LPTT tests, both linguistic and task fulfilment criteria are used in making judgements about performance. While the linguistic criteria focus on the components of language proficiency as traditionally conceived, such as accuracy, fluency and pronunciation, the task fulfilment criteria require raters to address such questions as:

> - Was the style and tone of delivery appropriate for the classroom?
> - Did the candidate tailor his/her language in such a way as to make it intelligible to second language learners?
> - Were the classroom instructions issued in a clear and convincing manner?
>
> (Elder, 1995)

Similarly, on the ClAsS (see Appendix 1) there are items on the schedule which pertain to aspects of 'teacherly' behaviour such as:

- clearly marks transitions from one idea/lesson stage to the next using words such as *so*, *now*, *right*, *we're going to*;
- deals effectively with wrong answers, nonresponse; e.g. by rephrasing questions/reviewing steps in a process.

Raters are required to consider these as well as purely linguistic features of performance in making overall judgements about communicative effectiveness. While all these features are undoubtedly language-based, they may draw on other aspects of 'teacherly' behaviour.

Combining these different dimensions of communicative competence can, however, be problematic. The analysis of rating patterns on the LPTT (Italian) showed that assessments made against the linguistic criteria were sometimes at odds with the ratings assigned for task fulfilment. A Rasch analysis of test-scores yielded misfitting ability estimates for 10 of the 75 candidates who sat for the speaking test, suggesting that the test as it stands, may be unworkable as a measurement procedure. Scrutiny of individual score profiles showed that the 'misfitting' candidates were those who achieved either a high score on the linguistic criteria and a low score for task fulfilment or vice versa (Elder, 1995).

In an attempt to find out the possible source of these 'disorderly' measurements, transcriptions of test discourse (recorded on videotape) were undertaken for each of the misfitting candidates, i.e., those who performed at consistently high levels on task fulfilment criteria and low for linguistic competence, on the one hand, and for those who scored high on the linguistic criteria and low for task fulfilment, on the other.

A short segment of performance from a representative of each of these two groups of candidates is set out below. The segments 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 perform a simple classroom construction activity – in this case, how to make a paper model of a sheep.

In describing one step of the construction activity (making a paper animal) Candidate A (who scored 'high' on task fulfilment and 'low' on linguistic competence) demonstrates what has to be done by demonstrating with her hands the action of curling a strip of paper with a pair of scissors and saying:

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

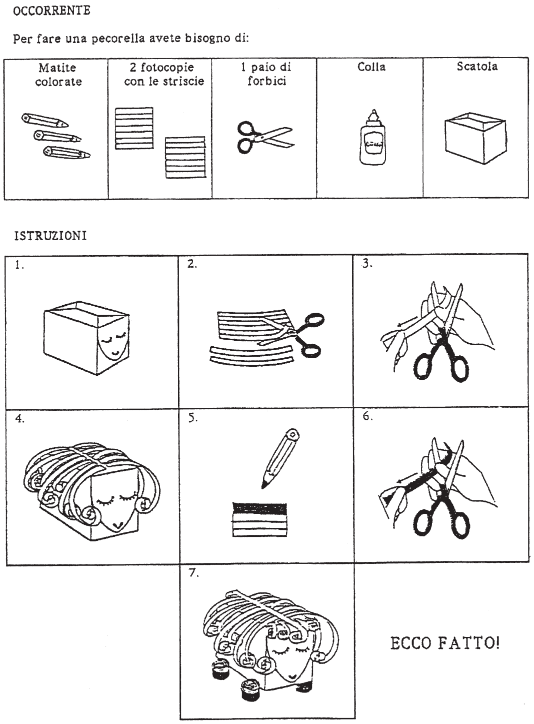


Figure 2 Sample task from LPTT (Italian) (Elder, 1995)

Her speech is delivered at a slower place and she repeats certain words.

Candidate B (scoring ‘high’ on linguistic competence and ‘low’ on task fulfilment), on the other hand, describes the action with words rather than gestures, using more sophisticated syntax and more precise lexis:

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

Note also that Candidate B (who is, in fact, a native speaker of Italian) chooses to use the more formal, impersonal *si* form of the imperative, whereas Candidate A opts for the simpler and more inclusive second person plural form.

Even from this small segment of the transcript 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 task fulfilment and why the opposite was true for the first candidate. Although Candidate A’s use of *così* (‘like this’) accompanied by a gesture may be a direct result of her lack of linguistic competence (i.e., she may not know the appropriate vocabulary to describe this step in the activity), she has resorted to a simplification strategy which is arguably more appropriate for young second learners with limited linguistic proficiency than is the more complex and linguistically precise utterance produced by Candidate B.

The above example (and there are others) draws attention to what may be a fundamental incompatibility between the traditional notion of general proficiency in a context which assumes a developmental continuum involving an increase in range and complexity of language use across all contexts of language use, and the nature of performance in specific situations where features of strategic competence 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 native speakers and other 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 nevertheless respond (whether consciously or unconsciously) more appropriately to the specific demands of the criterion domain. Widdowson (in press) makes a similar point:

> There is no way either of knowing how representative the learner’s performance is of a more general ability to communicate. The learner might be successful by the ingenious use of avoidance strategies, and these may not be distinguishable from an adherence to the least effort principle that characterises

> normal pragmatic uses of language. In this case, in effect, all you get is evidence of the so-called ‘strategic competence’ without knowing whether it is compensatory or not, or if it is, what it is compensating for.

On the CLAsS this incompatibility between linguistic and non-linguistic components of the test construct showed up in a lack of agreement between two groups of raters charged with the assessment of candidate ability: the subject specialists, naive raters in so far as they had had no language training, and the ESL teachers, who were accustomed to dealing with language difficulties but were not necessarily familiar with the content of the mathematics or science curriculum. For example, there was an unacceptably low correlation between the ratings assigned by the two groups in their evaluation of Item 2 on the observation schedule (Using subject-specific language; see Appendix 1). Feedback from the raters suggests that the reason for this disagreement is that the ESL teachers were focusing on the lexis, grammar and internal cohesion of the candidate’s classroom presentation and the pronunciation of specialist terminology, while the subject specialists were more concerned about the way in which subject content was conceptualized. These different orientations also affected each rater group’s final estimates of candidates’ overall communicative effectiveness, with disagreement as to whether the performance was satisfactory or unsatisfactory occurring in 20% of cases (Elder, 1993b).

The construct of teacher proficiency, as operationalized in these performance-based measures of teacher proficiency, is clearly multidimensional, and this poses problems for the interpretation and reporting of performance. One solution to this problem would be to separate the purely linguistic and the more classroom-specific aspects of performance in our reporting. In the case of LPTT we could set cut-offs on the basis of linguistic scores alone, but use information about classroom competence, as reflected in ratings assigned for task fulfilment, to assist in decisions about borderline cases where evidence of context sensitivity may serve to compensate for linguistic shortcomings. Moreover, in the CLAsS we should perhaps entrust the task of rating to an ESL teacher rather than a subject specialist. However, in theoretical terms this amounts to a weakening of the tests’ claim to specificity. If information about general proficiency is enough, and if the opinion of subject specialists doesn’t count, then there seems to be little point, other than to satisfy the need for face validity, in trying to capture the context-specific nature of language performance.

V Conclusion

This article has illustrated the problems associated with characterizing teacher discourse as a specific purpose domain (the problem of specificity), in attempting to simulate classroom-like behaviours in test situations (the problem of authenticity) and/or to assess classroom-related aspects of performance alongside other purely linguistic features (the problem of inseparability). Douglas (in press) sees these thorny issues as the basis for theory-building in LSP, suggesting that further research is needed in all three areas and implying that our understanding of what it takes to predict performance in specific domains of language use is still quite limited. At the same time, while the profession now takes for granted the value of assessing language in meaningful 'life-like' contexts, recent studies by scholars working in performance testing seem to give pause to that development. For example, Iwashita *et al*. (forthcoming), Norris *et al*. (1998), Papajohn (1999) and Freedle and Kostin (1999) all appear to be less concerned with the relationship between test-tasks and their real world counterparts than with intra-task effects on test-taker performance. In other words, they raise the question of whether changing test-task characteristics (prompts, topics, texts) or conditions (planning time, speededness), which are believed to increase or decrease their level of cognitive demand, will be reflected in test-scores assigned to candidates. This research is, I believe, indicative of both a growing concern about the indeterminacy of performance-based tasks as a means of measurement and a realization that the LSP testing enterprise of the 1980s and 1990s, in spite of its laudable attempt to capture the particularities of real world communication, raises more questions than it answers. As Skehan (1998) reminds us:

> There is still no way of relating underlying abilities to performance and processing conditions, nor is there any systematic basis for examining the language demands of a range of different contexts. As a result, it is not clear how different patterns of underlying abilities may be more effective in some circumstances than others, nor how these underlying abilities are mobilized into actual performance (Skehan, 1998: 159).

VI References

Bachman, L.F. 1991: What does language testing have to offer? *TESOL Quarterly* 25(4), 671–704.

Bachman L.F. and **Palmer, A.S.** 1996: *Language testing in practice*. Oxford: Oxford University Press.

Douglas, D. 2000: *Assessing languages for specific purposes: theory and practice*. Cambridge: Cambridge University Press.

—— in press: Three problems in testing language for specific purposes:

authenticity, specificity, and inseparability. In Elder, C., Brown, A., Hill, K.N., Iwashita, N., Lumley, T., McNamara, T.F., O'Loughlin, K., editors, *Experimenting with uncertainty: essays in honour of Alan Davies*. Cambridge: Cambridge University Press.

Douglas, D. and **Selinker, L.** 1993: Performance on a general versus a field-specific test. In Douglas, D. and Chapelle, C., editors, *A new decade of language testing research*. Selected Papers from the 1990 Language Testing Research Colloquium. Alexandria, VA: TESOL.

Elder, C. 1993a: Language proficiency as predictor of performance in teacher education. *Melbourne Papers in Language Testing* 2(1), 1–17.

—— 1993b: How do subject specialists construe classroom language proficiency? *Language Testing* 10(3), 235–54.

—— 1993c: *The proficiency test for language teachers: Italian, Volume 1: Final report on the test development process*. Melbourne: NLLIA Language Testing Centre, University of Melbourne.

—— 1994: Performance testing as benchmark for foreign language teacher education. *Babel. Journal of the Federation of Modern Language Teachers Associations* 29(2), 9–19.

—— 1995: Are raters' judgements of language teacher effectiveness wholly language-based? *Melbourne Papers in Language Testing* 3(2), 40–59.

Elder, C. Iwashita, N. and **Brown, A.** 1995: *The proficiency test for language teachers: Japanese, Volume 1: Final report on the test development process*. Melbourne: NLLIA Language Testing Research Centre, University of Melbourne.

Freedle, R. and **Kostin, I.** 1999: Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing* 16(1), 2–32.

Grant, L. 1997: Testing the language proficiency of bilingual teachers: Arizona's Spanish proficiency test. *Language Testing* 14(1), 23–46.

Hinofotis, F., Bailey, K. and **Stern, S.** 1981: Assessing the oral proficiency of prospective FTAs: Instrument development. In Palmer, A.S., Groot, P.J.M. and Trosper G., editors, *The construct validation of tests of communicative competence*. (pp. 106–126) Washington, DC, Teachers of English to Speakers of Other Languages.

Iwashita, N., McNamara, T. and **Elder, C.** forthcoming: Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information processing approach to task design. *Language Learning*.

Jacoby S. and **McNamara, T.** 1999: Locating competence. *English for Specific Purposes* 18(3), 213–41.

Jones, R. 1979: Performance testing of second language proficiency. In Brière, E. and Hinofotis, F., editors, *Concepts in language testing: some recent studies*. Washington DC: TESOL, 49–57.

Lumley, T. and **Brown, A.** 1996: Specific purpose language performance tests: task and interaction. In Wigglesworth, G. and Elder, C., editors, *The language testing cycle: from inception to washback*. Australian Review of Applied Linguistics, Series S, Number 13, 105–36.

McNamara, T.F. 1996: *Measuring second language performance*. London and New York: Addison Wesley Longman.

Messick, S. 1993: Validity. In Linn, R.L., editor, *Educational measurement* (3rd edn). National Council on measurement in Education and American Council on Education, Oryx Press, 13–304.

Nicholas, H. 1993: Languages at the crossroads; the report of the national enquiry into the employment and supply of teachers of languages other than English. Victoria: National Language and Literacy Institute of Australia.

Norris, J., Brown, J.D., Hudson, T. and **Yoshioka, J.** 1998: *Designing second language performance assessments.* Second Language Teaching & Curriculum Center, University of Hawai'i at Manoa.

Papajohn, D. 1999: The effect of topic variation in performance testing: the case of the chemistry TEACH test for international teaching assistants. *Language Testing* 16(1), 52–81.

Plakans, B. and **Abraham, R.** 1990: The testing and evaluation of international teaching assistants. In Douglas, D., editor, *English language testing in U.S. colleges and universities*. Washington, DC: NAFSA.

Skehan, P. 1998: *A cognitive approach to language learning*. Oxford: Oxford University Press.

Smith, J.A. 1992: Topic and variation in oral proficiency in international teaching assistants. PhD dissertation, University of Minnesota DA 9217683.

Stevenson, D.K. 1985: Authenticity, validity and a teaparty. *Language Testing* 2(1), 41–47.

Swales, J. 1990: *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Teasdale, A. 1994: Authenticity, validity and task design for tests of well-defined ESP domains. In R. Khoo, editor, *The practice of LSP: perspectives, programmes and projects*. Singapore: SEAMEO Regional Language Centre, 230–24.

Viete, R. 1998: Culturally sensitive and equitable assessment of oral English for overseas-qualified teacher trainees. *Journal of Intercultural Studies* 19(2), 171–84.

Wesche, M. 1992: Second language performance testing: the Ontario test of ESL as an example. *Language Testing* 4, 28–84.

Widdowson, H. in press: Communicative language testing: the art of the possible. In Elder, C., Brown, A., Hill, K.N., Iwashita, N., Lumley, T., McNamara, T.F., O'Loughlin, K., editors, *Experimenting with uncertainty: essays in honour of Alan Davies*. Cambridge: Cambridge University Press.

Appendix 1 Classroom language assessment schedule (Elder, 1993b: 251–53)

1. General language proficiency

Rating for *Intelligiblity* of expression	highly satisfactory	acceptable	at risk	unsatisfactory

		Comment (*strengths & weaknesses*)	Needs work
1.1.	projects and pitches voice appropriately		□
1.2.	pronounces words/sounds clearly		□
1.3.	utters sentences clearly (i.e. with suitable rhythm and intonation)		□
1.4.	clearly distinguishes questions, statements and instructions		□
1.5.	stresses important words/ideas (says them louder, more slowly, with pauses)		□
1.6.	clearly marks transitions from one idea/lesson stage to the next using words such as *so*, *now*, *right*, *we're going to*		□
1.7.	uses appropriate facial expression gesture, body movement		□

Rating for *Fluency & flexibility* of expression	highly satisfactory	acceptable	at risk	unsatisfactory

1.8.	speaks at appropriate speed		□
1.9.	speaks fluently (i.e., not too much stumbling hesitation, groping for words)		□
1.10.	can express ideas in different ways (e.g. by rephrasing, elaborating, summarizing)		□

Rating for *Accuracy* of expression	highly satisfactory	acceptable	at risk	unsatisfactory

		Comment (*strengths & weaknesses*)	Needs work
1.11.	grammar of spoken and written English is generally accurate		□
1.12.	formulates questions clearly		□
1.13.	uses correct spelling and punctuation in boardwork and handouts		□

Rating for *Comprehension* of expression	highly satisfactory	acceptable	at risk	unsatisfactory

		Comment (*strengths & weaknesses*)	Needs work
1.14.	demonstrates understanding of student language		□
1.15.	seeks clarification of student language when necessary (e.g. asks them to repeat/rephrase)		□

2. Using subject-specific language

Rating for *Subject-specific* language	highly satisfactory	acceptable	at risk	unsatisfactory

		Comment (*strengths & weaknesses*)	Needs work
2.1.	demonstrates knowledge of scientific and mathematical terms		□
2.2.	pronounces specialist terms clearly		□
2.3.	uses specialist terms judiciously (grading them and writing them on the board when appropriate)		□
2.4.	makes clear the connections between ideas (stress link words *if*, *since*, *in order*)		□
2.5.	explains scientific and mathematical processes/concepts in ways appropriate to the audience (using simple language, familiar/concrete examples)		□
2.6.	explains diagrams/models/use of equipment clearly		□
2.7.	description/definition of terms/processes is a usable model for students' written assignments		□

3. Using the language of classroom interaction

Rating for language *Classroom interaction*	highly satisfactory	acceptable	at risk	unsatisfactory

		Comment (*strengths & weaknesses*)	Needs work
Involvement of students in class and lesson content			
3.1.	uses variety of forms of address (we, you, us/student names)		□
3.2.	poses questions to check understanding of previously learned material/new information		□
3.3.	grades questions appropriately for students and learning task: simpler to more complex; closed/open		□
3.4.	offers questions to individuals and whole class		□
3.5.	clearly signals acceptance/rejection of student response		□
3.6.	responds appropriately to students' questions, requests for assistance		□
3.7.	deals effectively with wrong answers, non-response (e.g. by rephrasing questions/reviewing steps in a process)		□
3.8.	adopts appropriate level of formality/firmness		□
3.9.	gives clear instructions		□
3.10.	maintains contact with class while dealing with individual demands/using blackboard, etc.		□

Overall communicative effectiveness

Rating for *Overall effectiveness*	highly satisfactory	acceptable	at risk	unsatisfactory

Appendix 2 Language proficiency test for teachers (Italian): brief test description

Phase	Task	Audience	Mode
1. Warm-up and discussion	Brief getting to know you conversation and discussion of aspects of teacher role	Interviewer	Dialogue
2. Read aloud	Read aloud a short story and explain meaning of selected words from the passage	Whole class (imagined)	Monologue
3. Story retell	Retell the same story using picture book prompts	Whole class (imagined)	Monologue
4. Instructions	Give directions as to how to carry out a simple construction activity using picture prompts	Whole class (imagined)	Monologue
5. Assign and model a roleplay	Explain participant roles using cue cards and perform roleplay	Interviewer (as student)	Dialogue
6. Explain learner error	Explain specified mistakes in student writing	Interviewer (as student)	Dialogue