Language for Specific Purposes assessment criteria: where do they come from?

Dan Douglas *Iowa State University*

Typically in assessment of Language for Specific Purposes (LSP), test content and methods are derived from an analysis of the target language use (TLU) situation. However, the criteria by which performances are judged are seldom derived from the same source. In this article, I argue that LSP assessment criteria should be derived from an analysis of the TLU situation, using the concept of indigenous assessment criteria (Jacoby, 1998). These criteria are defined as those used by subject specialists in assessing communicative performances of both novices and colleagues in academic, professional and vocational fields. Performance assessment practices are part of any professional culture, from formal, gatekeeping examination procedures, to informal, ongoing evaluation built into everyday interaction. I suggest a procedure for deriving assessment criteria from an analysis of the TLU situation and explore problems associated with doing so, recommending a 'weak' indigenous assessment hypothesis to assist in the development of LSP test assessment criteria and guide interpretations of test performance.

I Introduction

A colleague of mine at Iowa State University, Rebecca Burnett, who teaches a doctoral course in professional communication recounted to me not long ago how, in a class in which she taught students to use technical register, populated mainly by native speakers of English, two international students – one from Ukraine and one from Russia – were often dismayed when she returned their papers to them. They assumed that the 'errors' that Burnett had commented on in the papers were due mainly to their faulty English language skills. Burnett assured them that this was not the case: the great majority of the features she had marked were no different from those found in the papers of the native English-speaking students: their problems were not with English so much as with the register of technical communication.

My colleague's story resonated with a point made by Sally Jacoby in her dissertation (Jacoby, 1998), a study of conference presentation

Address for correspondence: Dan Douglas, TESL/Applied Linguistics Program, English Department, Iowa State University, Ames, IA 50011, USA; e-mail: dandoug@iastate.edu

Language Testing 2001 18 (2) 171–185 0265-5322(01)LT200OA

rehearsals among physicists. Among her findings, Jacoby discovered that the physicists, in assessing each others' practice presentations, generally applied their assessment criteria to all presenters, without regard to their status as native or nonnative speakers of English. Moreover, Jacoby and McNamara (1999) point out that not only were the same criteria applied to all presenters, but that the standards of excellence were the same for all: no normative standard based on the notion of native speaker, including those of linguistic accuracy and style, was in force.

It would appear that specific purpose language performance assessment criteria devised by second language testing/teaching professionals may be different from those of most concern to nontesting/teaching professionals. In this article I explore the question of where assessment criteria come from in testing in Language for Specific Purposes (LSP), and how we might work towards a procedure for deriving them from the same source that we derive LSP test content and methods: from an analysis of specific purpose language use situations.

II Background

In traditional general purpose language testing, it is usually the case that test content is derived from a theory of language ability, such as that outlined by Bachman (1990), a theory of language acquisition, such as that proposed by Pienemann *et al*. (1988) or a course syllabus which itself is based on a theory of language ability or acquisition. Furthermore, in general purpose testing, test methods are usually derived from psychometric theories about how best to measure cognitive constructs such as communicative language ability. This theoretical orientation in general purpose testing is a necessary consequence of the fact that the situations in which the language being tested will eventually be used are not specifiable in any great detail because they are largely unknown. Widdowson (1983) has suggested that the goal of general English teaching is to develop 'communicative capacity' in learners that will equip them to achieve diverse communicative goals after completing the course. It is my view that general purpose language tests are thus designed to measure this communicative capacity without substantial reference to the situations in which the language will be used, beyond 'situational window dressing' or, as it is known in the assessment field, face validity.

LSP tests, on the other hand, derive their content from an analysis of specific language use situations of importance to the test-takers. True, the analysis is guided by theoretical frameworks, but

the point is that LSP test-developers can and do find out in detail during the test development process what situations the test-takers will find themselves in and are able to draw on the linguistic and situational features to obtain the material for test development (McNamara, 1997). Interestingly, though, too, in LSP testing, the test methods themselves may also be derived from the analysis of the target situation. The tasks that language users typically perform in the target situation can be translated into test tasks by reference to task features, such as those proposed by Bachman and Palmer (1996). This aspect of LSP testing is most clear in performance tests, such as the *TEACH*, a performance test for international instructors at my university in which the candidate presents a short lesson in his or her field to a small group of undergraduate students and responds to questions from them as would happen in a nontest university classroom situation (Abraham and Plakans, 1988). Another example is the Proficiency Test in English Language for Air Traffic Controllers (Institute of Air Navigation Services, 1994) in which test-takers have to listen through earphones to messages from pilots and respond to them appropriately, as they would do in the actual situation.

The procedures for analysing the target language use (TLU) situation in terms of the features of the specific purpose language and tasks that provide the content and methods for LSP tests are fairly well understood (e.g., Bachman and Palmer, 1996; Douglas, 2000), if not universally applied. However, the derivation of the criteria by which we judge performance on our tests is a different matter.

III The derivation of assessment criteria

Assessment criteria, in both general and specific purpose testing, have traditionally been derived from the same theories of language knowledge and psychometrics. For example, McNamara (1996: 19) points out that 'These criteria often make implicit reference to a psychological construct or constructs which then emerge as the object of measurement'. There is very little discussion in the standard language testing literature about the provenance of assessment criteria (but see Alderson *et al*., 1995; Bachman and Palmer, 1996). Discussion usually focuses on accuracy and consistency in applying the criteria. This is not to say that rating criteria derived from the TLU situation do not overlap with the more theoretically derived criteria proposed by, for example, Bachman and Palmer (1996), but I wish to suggest that the target situation can provide insights into other types of criteria

that may be of importance to practitioners in those fields and which are not necessarily evident to language testing professionals. The point I want to emphasize is that, contrary to the cases of LSP test content and method, LSP assessment criteria have not usually been derived from an analysis of the TLU situation. Rather, they tend to have come from theoretical understandings of what it means to know and use a language (Jacoby and McNamara, 1999), without regard, in some cases, for the situation in which it is used (see, for example, Bachman and Palmer, 1996). There are exceptions, which I will discuss below – two teaching performance tests and one for trainee tour guides – in which assessment criteria were derived from analyses of TLU situations. However, before looking at these examples, I want to discuss the theoretical underpinnings of the development of LSP assessment criteria and argue that it is important for us to derive them not only from a theoretical understanding of communicative language ability (Bachman and Palmer, 1996) but also from an empirical analysis of the TLU situation.

In an article that will become essential reading for LSP testers, Jacoby and McNamara (1999) discuss the 'primarily linguistic orientation' of LSP assessment, with particular reference to the Occupational English Test (OET), an Australian performance-based test of English for immigrant and refugee health professionals. They compare the assessment of medical communication skills among native English-speaking health professionals with that of the immigrant candidates and find that, whereas the communication skills of the native English-speaking medical undergraduates is assessed along with other aspects of medical competence, the language skills of immigrants and refugees are assessed quite separately from clinical knowledge and skills. The separation of language skills from medical skills among the ESL medical population is due to legislation that mandates that the OET assess English ability but not medical competence, which is assessed in a second-stage procedure conducted by health professionals. This separation makes a certain amount of sense from the standpoint of policy, but the consequences of attempting to assess language ability for specific health-related purposes using a set of linguistically derived rating criteria that make no reference to the medical context of the test content and methods may be problematic.

The primary developer of the OET, McNamara (1996), reported that he based the scoring categories, or assessment criteria, he used in the speaking and writing subtests of the OET on the Foreign Service Institute (FSI) Oral Interview, developed in 1957, itself based on a comparison with the assumed language proficiency of

native speakers (Wilds, 1975). The OET speaking criteria thus, similar to those of the FSI, include overall communicative effectiveness, intelligibility, fluency, comprehension, appropriateness of language, and resources of grammar and expression. The case of the OET with respect to the derivation of assessment criteria is not at all unusual in language test development. However, McNamara (1996) reports that about six years after the introduction of the OET, test supervisors were receiving complaints both from physicians conducting the clinical examinations of the overseas candidates and from hospital supervisors of the overseas physicians working in medical practice that the English skills of those who had passed the OET appeared inadequate for interactions with both patients and medical colleagues. As a consequence, Lumley *et al.* (1994) carried out a series of investigations comparing the judgements of the physicians with those of the usual OET examiners, who had been charged with being too lenient. To everyone's surprise, there was no difference between the physicians and the trained examiners in terms of standards applied. Jacoby and McNamara conclude from this that 'whatever the doctors were complaining about was *not* being captured by the OET' (1999: 223). They note that the 'inevitable simplification and dilution of the real-world task when simulated in performance test conditions' calls into question the validity of the OET, and go on to speculate that 'it is also possible that the discrepancy between the test performance and reported real-world performance can be accounted for in terms of the criteria used to judge the performance . . .' (1999: 223–24).

IV Indigenous assessment criteria

Jacoby and McNamara then consider Jacoby's investigation of conference rehearsals by a group of physicists (Jacoby, 1998). In these rehearsals, each participant would present his or her paper to the rest of the group, under conditions similar to that of the conference, and the members of the group would provide feedback on the performance based on implicit criteria, which Jacoby calls 'indigenous assessment criteria'. She defines such criteria as those used by subject specialists in assessing the communicative performances of apprentices in academic and vocational fields. Jacoby and McNamara found some significant disparities between the criteria the physicists used to judge each others' language performances and those employed in the OET – for very good reasons, of course, given the different purposes and contexts of the two situations – but the comparison gives us, I think, some insights into

the nature of indigenous assessment criteria that will be useful in the development of LSP tests.

Performance assessment practices are part of any professional culture, from formal, gatekeeping examination procedures, to informal, ongoing evaluation built into everyday interaction with novices, colleagues and supervisors. Indeed, professional development is just a specialized form of socialization, a general process long recognized as the vehicle through which culturally specific language, discourse, cognition and skills are transmitted and developed through social interaction (for a review of the literature, see Jacoby, 1998). Experienced, competent professionals are able to articulate assessments, the criteria employed, and ways in which language performances might be improved to both colleagues and the persons being assessed. However, since the professionals normally externalize their criteria only in the context of authentic communicative situations in their work, the criteria are accessible to researchers primarily by means of an analysis of the discourse in which they are displayed. The researchers, therefore, need to engage in very careful study of the assessment interaction and discourse in the TLU situation, with help from discourse analysts and from specialists in the target field. Other examples of studies of indigenous assessment criteria include one by McNamara and colleagues who are studying the indigenous assessment criteria articulated by medical practitioners (McNamara, 1997) and that by Douglas and Myers, who studied the criteria used by veterinary professionals in assessing the communication skills of veterinary students (Douglas and Myers, 2000). The investigation of indigenous assessment is still a very new, undeveloped possibility for specific purpose language testing; however, the expectation is that the study of various types of assessment activities in professional and vocational settings will help test-developers to establish criteria for the specific purpose testing enterprise. I do not advocate throwing out theoretically-based approaches to the development of assessment criteria, but rather supplementing our understanding of the complex construct of communicative language use in specific purpose contexts by taking into account the criteria deemed important by experienced professionals in the various fields for which we produce tests. The goal is to make the assessment criteria guiding our interpretations of language test performance as congruent with the specific purpose situation as are the test content and methods. Jacoby and McNamara (1999) caution, however, that there are difficult problems associated with applying these indigenous criteria – derived from highly specific, dynamic contexts of use – to language tests, no matter how situationally authentic the

tests may be. I discuss some of the potential problems later in this article.

Examples of the types of indigenous assessment criteria that may emerge from an analysis of the TLU situation are shown in Table 1. It is important to note that such criteria as these may differ from target situation to target situation, but preliminary research has shown a surprising amount of overlap in the sorts of criteria various stakeholder groups may arrive at. Douglas and Myers (2000) found that veterinary professionals, a group of applied linguists, and veterinary students identified quite similar sets of criteria in evaluating videotaped veterinarian–client interviews, as shown in Table 2. The three groups differed in some respects in the criteria they employed in assessing the interview performances: neither the students nor the applied linguists identified 'rapport'; neither veterinarian group saw 'appropriacy' as relevant to making their judgements. However, 10 of 15 categories overlapped.

Two points may be emphasized with regard to indigenous criteria in the physics presentations and veterinary interviews. First, as I mentioned above, they are quite different from those used in the OET; however, the point is that, whereas in the OET the criteria are 'rooted in a comparison with a normative construction of native speaker', the indigenous criteria are 'rooted in the complex task of presenting an effective conference talk' (Jacoby and McNamara, 1999: 233). Rooted, in other words, in the TLU situation itself: 'inextricably intertwined with the content, argumentation structure, and multi-modality' of the physics presentation (Jacoby and McNamara, 1999: 234), in contrast to those of the OET which are derived from the general purpose FSI scale. Now, it may very

Table 1 Physics conference presentations assessment criteria

- overall quality of the performance;
- keeping to the time limit;
- articulating the significance of the topic to the profession;
- designing visuals to accompany the talk which are coherent and legible;
- stating arguments and labeling visuals clearly;
- stating information, arguments and rhetorical steps explicitly and completely;
- avoiding verbosity;
- making effective, convincing arguments;
- maintaining accuracy of content;
- delivering a technically polished performance (in terms of volume, rate, body positioning, management of the visuals);
- avoiding linguistic errors.

Source: Jacoby and McNamara, 1999: 229 ff.

Table 2 Summary of veterinary interview assessment criteria by three groups

Vet professionals	Vet students	Applied linguists
*Introduction** none of 'em had a very good introduction	*Introduction** I don't know if I introduced myself in that uh in that introduction	*Opening** he asked if he could call her Tricia
*Rapport** a good idea to establish more rapport with the client – more chit chat	(No exemplars)	(No exemplars)
*Demeanor** some of his questions would make some clientele rather defensive	(No exemplars)	*Authority/confidence** he had more authority at the end
*Knowledge base** he's not overly familiar with li- some of the livestock management terminology	*Knowledge** if I was really a doctor I should have known	*Knowledge** he should have known that they were – that they were not milk cows
Follow up/elicitation** really strong follow-up questions being able to elicit the information	*Follow-up/interviewing skills* I tried to get her to elaborate as much as she could on some major areas	*Follow-up*/getting information** he just should've followed up
*Phraseology** his wording of questions was not ideal	*Phraseology* uh that was kinda stuttered – staggered – it wasn't a kind of well phrased question	*The way he asks the questions** I don't think that's the way a vet would ask that question
Level didn't talk down to her at all	*Level** I think I'm doin' a good job talking on her level because that's the same level that I'm at	(No exemplars)
*Pace** taking more time between his questions	*Pace** I maybe could have slowed the pace down a little bit	*Duration** it seemd to me that that was much longer – that interview
Clarification made sure he understood what I was saying	*Clarification* okay – it's important to know whether a chronic or an acute type of process	*Clarification* he repeated some of her – some things she had said to make sure that he got it straight
*Structure** the interview has various components to it – it's got an overall structure	*Structure* maybe I would've prepared questions better	*Structure/cohesion** he doesn't give any framework t' her

Table 2 Continued

Vet professionals	Vet students	Applied linguists
*Coverage/depth/breadth** he just focused completely in on one thing	*Coverage* I think I did a decent job of turning over most of the stones that coulda been turned over	*Content* but the content – he didn't ask about the water
*Appearance** they all looked clean 'n well groomed	(No exemplars)	(No exemplars)
(No exemplars)	(No exemplars)	*Appropriacy** inappropriate responses
*Engagement** really engages the person	*Engagement* I'm tryin' to maintain pretty good eye contact with her	*Engagement** he doesn't look at all engaged
*Summary** I don't think he gave an overall summary	*Summary* tried to sum up what she said	*Summary** right at the end he summed it up

Note: *Name provided by participants.
Source: Douglas and Myers, 2000.

well be that language testers have learned a lot about LSP assessment criteria since the original development of the OET, and that few would use a general purpose scale in a specific purpose test, yet Douglas (2000), in reviewing a number of specific purpose language tests, found that many recent examples continue to employ rather traditional, linguistically-oriented criteria. Secondly, in the physics indigenous criteria, no distinction is made between native and nonnative speakers of English, with the exception of the category of 'linguistic error', which is directed exclusively towards the nonnative English speakers, and then only when errors appear in the visuals. This is in contrast to the OET criteria, which are based on an implicit comparison with native speaker performance embodied in the FSI scale. Jacoby and McNamara (1999: 234) observe that the OET criteria are 'oddly out of synch' with the long-held LSP principle that 'special purpose performance is by definition task-related, context-related, specific, and local'.

This echoes the point that I made at the outset that, while LSP tests derive their content and methods from analyses of specific purpose language use situations, typically the criteria by which LSP performances are assessed are derived from theoretical understandings of language knowledge and use without regard for context of situation. At the very least an analysis of the indigenous

assessment criteria in the specific purpose domain in which we are attempting to develop a language test could serve as a framework for the development of assessment criteria in the test domain. As Jacoby and McNamara note, it is not a simple matter to make the transition from the TLU indigenous criteria to the criteria that will be employed in the test: Where 'indigenous assessment is a here-and-now interactional problem solving activity' (1999: 235), language testers are, by the nature of their activity, looking for generalizability across language performances – in McNamara's 'weak' sense of a language performance test (McNamara, 1996) – as well as taking account of such psychometric qualities as reliability and validity. Yet, we must, I think, seek to complement the important features of language testing practice with assessment practices in the real world contexts of interest to LSP testers.

V Some current examples

There are some exemplary tests currently in use in which the test-developers investigated indigenous assessment criteria (although they did not use the term) as part of the test development process. In the *TEACH* test for international instructors mentioned above, assessment criteria were derived from an analysis of classroom performances, supplemented by undergraduate students' perceptions of teacher quality, and include such characteristics as familiarity with cultural code, appropriate nonverbal behaviour, rapport with class, development of explanation, clarity of expression, use of supporting evidence, eye contact, use of chalkboard, and teacher presence. These are reminiscent in quality of some of those derived from the physics presentation rehearsals.

A second example, another test for teachers, the *Proficiency test for language teachers: Italian* (Elder, 1993b), was developed by staff at the National Language and Literacy Institute of Australia, Language Testing Research Centre at the University of Melbourne in 1993. The test has two main functions: (1) to serve as a benchmark for the language requirements of the foreign language teacher and (2) to certify language teachers for employment in primary schools. The test-developers observed Italian language teachers in action in three primary schools and one junior secondary school; they also used the language teachers as specialist informants, consulting them about curriculum and textbooks and classroom procedures (Elder, 1993b). The job analysis phase of the development process helped determine the assessment criteria, which include what Elder calls 'classroom communicative competence' criteria: 'teacherliness', the quality of language production in terms of its

suitability for the classroom; and 'metalanguage', the quality of the test-taker's explanations of learner error. Thus, the assessment criteria are grounded in the context of the foreign language classroom, the TLU situation.

A third example was also produced by the Language Testing Research Centre at the University of Melbourne: the *Japanese language test for tour guides* (1992). It has the dual purpose of indicating to employers the language proficiency of applicants for positions as Japanese-speaking tour guides, and as a selection criterion for applicants to tour-guide training courses (Brown, 1995). The 30-minute test, consisting of five role plays, was produced in consultation with experienced tour guides, and the assessment criteria were based on their judgements about the necessary features of quality tour-guide communication, referred to as 'task fulfilment criteria' (Brown, 1993), which include such features as enthusiasm, empathy, making something sound interesting, persuasiveness, and awareness of interlocutors' needs or desires. Thus, again we see a test in which the assessment criteria were drawn, at least in part, from an analysis of indigenous criteria in the TLU situation.

VI Procedures for analysing TLU situations for assessment criteria

A basic procedure that LSP test-developers can draw upon in the investigation and description of indigenous assessment criteria in the TLU situation is 'grounded ethnography in context based research'. Grounded ethnography (Frankel and Beckman, 1982) is an approach to describing and understanding a TLU situation from the perspective of language users in that situation. The technique has been defined as: 'a means for the researcher to understand an event by studying both its natural occurrence and the accounts and descriptions of it provided by its coparticipants' (Frankel and Beckman, 1982: 1).

Ethnography itself is, of course, an approach to the study of behaviour from the differing viewpoints of the participants; it has been in use since the late 1960s. What Frankel and Beckman bring to ethnographic research is the concept of 'grounding', in which the viewpoints of the participants are derived from their own observation of videotaped recordings of the events under analysis. Douglas and Selinker (1994), building on this work in grounded ethnography, have provided a number of guidelines for 'context-based' research: the study of second language acquisition and use

in important real-life contexts. Douglas and Selinker make a distinction between primary data and secondary data:

- primary data: the interlanguage talk or writing we wish to study.
- secondary data: commentary on the primary data (Douglas and Selinker, 1994: 120).

Douglas and Selinker discuss two categories of secondary data: commentaries on the primary data by the participants themselves and various types of expert commentaries upon the primary data. As sources of this type of secondary data, we have worked with other linguists, ethnographers and ethnomethodologists, and specialist in the target fields, each of whom bring their various perspectives and methods to bear on the primary data.

In addition, it is essential in context-based research that the LSP tester make use of what Selinker (1979) has called 'subject specialist informant procedures' in a principled way in analysing the TLU situation in LSP disciplines in which the test-developers have little or no expertise. In addition to Selinker's own work, there is some preliminary research (Elder, 1993a) which suggests that such informants are capable of playing a role not only in analysis of the situation, but also in the assessment of specific purpose language ability. Elder suggests that subject specialist raters of communicative ability may take a strong view of communicative performance, judging communicative success rather than the quality of language, *per se*. Elder's work also suggests that, while there is a substantial relationship between subject specialists' assessment of overall ability and that of language specialists, subject specialist judges do, in fact, sometimes assess specific purpose language differently from language specialists. Test-developers would do well to attend to the criteria they employ.

As for deriving the assessment criteria from the data, Jacoby (personal communication 1996) suggests that the commentary from all the informants be transcribed and submitted to an analysis:

1) The secondary data is analysed for the various comments raised or alluded to by group members, for the criteria they mention in relation to specific behaviours observed in the primary data; and
2) The list of specific comments is collapsed into a smaller, more generalized list of assessment criteria.

As Douglas (2000) has pointed out, the second step is problematic, since comments tend to focus on each particular performance and are highly context embedded. I have suggested reference to

the construct definition in the LSP test as a way of achieving 'a more generalizable set of correctness criteria since the criteria for correctness will reflect both the characteristics of specific purpose language ability in the target situation and the constraints placed on what aspects of the ability will be measured' (Douglas, 2000: 70). Sometimes informants will provide names of domains of talk they are engaged in or which are important to them (Douglas and Selinker, 1994), and to the extent possible in making the more general lists, terminology suggested by the informants themselves in the course of their discussions should be employed, making the task of categorizing easier.

VII Conclusion

What I want to emphasize in conclusion is the importance of considering assessment criteria that are derived from the analysis of the TLU domain in the development of LSP tests. The interpretations we make of test-takers' performances on our LSP tests will stand a much better chance of being appropriate in the specific purpose context as perceived by subject specialists if they are grounded in assessment criteria derived from an analysis of the TLU domain. There is a 'strong' indigenous assessment hypothesis which would involve employing the criteria derived from an analysis of assessment practices directly in the TLU situation; however, I do not advocate such a strong case. Rather, I wish to suggest a weaker indigenous assessment hypothesis in which the indigenous criteria may be used first to supplement linguistically-oriented criteria in line with the construct definition, and, secondly, to help guide our interpretations of language performances in specific purpose tests. Although there is a need to balance psychological/psychometric considerations with indigenous criteria, it seems to me that just as we mine the TLU situation for LSP test content and methods, there is much to be gained from going to that same source for assessment criteria. This is because, in LSP testing, that is where they properly come from.

VIII References

Abraham, R. and **Plakans, B.** 1988: Evaluating a screening/training program for nonnative speaking teaching assistants. *TESOL Quarterly* 22, 505–08.

Alderson, J.C., Clapham, C. and **Wall, D.** 1995: *Language test construction and evaluation.* Cambridge: Cambridge University Press.

Bachman, L. 1990: *Fundamental considerations in language testing.* Oxford: Oxford University Press.

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

Brown, A. 1993: The role of test-taker feedback in the test development process: test-takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing* 10(3), 277–303.

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

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

Douglas, D. and **Myers, R.K.** 2000: Assessing the communication skills of veterinary students: whose criteria? In Kunnan, A., editor, *Fairness in validation in language assessment.* Selected papers from the 19th Language Testing Research Colloquium. Studies in Language Testing 9. Cambridge: Cambridge University Press, 60–81.

Douglas, D. and **Selinker, L.** 1994: Research methodology in context-based second language research. Chapter 6 in Tarone, E., Gass, S. and Cohen, A., editors, *Methodologies for eliciting and analyzing language in context.* Northvale, NJ: Erlbaum, 119–31.

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

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

Frankel, R. and **Beckman, H.** 1982: IMPACT: an interaction-based method for preserving and analyzing clinical transactions. In Pettigrew, L., editor, *Explorations in provider and patient transactions*. Memphis, TN: Humana, 71–85.

Institute of Air Navigation Services. 1994: *PELA: A test in the proficiency in English language for air traffic control*. Luxembourg: Institute of Air Navigation Services.

—— 1998: *Science as performance: socializing scientific discourse through conference talk rehearsals*. Unpublished doctoral dissertation, University of California, Los Angeles.

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

Lumley, T., Lynch, B. and **McNamara, T.** 1994: A new approach to standard setting in language assessment. *Melbourne Papers in Language Testing* 3(2), 19–40.

McNamara, T. 1996: *Measuring second language performance*. London: Longman.

—— 1997: Performance testing. In Clapham, C. and Corson, D., editors, *Encyclopedia of language and education, Volume 7: Language Testing and Assessment*. Dordrecht, NL: Kluwer Academic Publishers, 131–39.

Pienemann, M., Johnson, M. and **Brindley, G.** 1988: Constructing an acquisition-based procedure for second language assessment. *Studies in Second Language Acquisition* 10, 217–43.

Selinker, L. 1979: On the use of informants in discourse analysis and language for specific purposes. *International Review of Applied Linguistics* 17, 189–215.

Widdowson, H. 1983: *Learning purpose and language use*. Oxford: Oxford University Press.

Wilds, C. 1975: The oral interview test. In Jones, R. and Spolsky, B., editors, *Testing Language Proficiency*. Arlington, VA: Center for Applied Linguistics, 29–44.