Assessing Language for Specific Purposes

Dan Douglas
# Contents

*Series editors’ preface*  
*Acknowledgements*  

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  Why test language for specific purposes?</td>
<td>1</td>
</tr>
<tr>
<td>2  Specific purpose language ability</td>
<td>24</td>
</tr>
<tr>
<td>3  Context, discourse domains, and task characteristics</td>
<td>41</td>
</tr>
<tr>
<td>4  Strategic competence: between knowledge and context</td>
<td>75</td>
</tr>
<tr>
<td>5  From target language use to test tasks</td>
<td>92</td>
</tr>
<tr>
<td>6  Specific purpose tests of listening and speaking</td>
<td>129</td>
</tr>
<tr>
<td>7  Specific purpose tests of reading and writing</td>
<td>189</td>
</tr>
<tr>
<td>8  LSP test development and technology</td>
<td>246</td>
</tr>
</tbody>
</table>

*Appendix: other LSP tests*  
*References*  
*Index*
CHAPTER ONE

Why test language for specific purposes?

Introduction

Testing language for specific purposes (LSP) refers to that branch of language testing in which the test content and test methods are derived from an analysis of a specific language use situation, such as Spanish for Business, Japanese for Tour Guides, Italian for Language Teachers, or English for Air Traffic Control. LSP tests are usually contrasted with general purpose language tests, in which purpose is more broadly defined, as in the Test of English as a Foreign Language (TOEFL) (Educational Testing Service 1965). As you will see, it is important to note that tests are not either general purpose or specific purpose – all tests are developed for some purpose – but that there is a continuum of specificity from very general to very specific, and a given test may fall at any point on the continuum. I will argue later in this chapter that LSP testing is a special case of communicative language testing, since both are based on a theoretical construct of contextualized communicative language ability, and that LSP tests are no different in terms of the qualities of good testing practice from other types of language tests.

I should note that, over the years since its beginnings, specific purpose language testing has been criticized on a number of grounds: specific purpose language proficiency is really just general purpose language proficiency with technical vocabulary thrown in; we don’t need specific purpose tests since, if we test general language knowledge, specific uses will take care of themselves; specific purpose
language tests are unreliable and invalid since subject knowledge interferes with the measurement of language knowledge; there is no theoretical justification for specific purpose language testing; and specific purpose language testing is impossible anyway, since the logical end of specificity is a test for one person at one point in time. In this book, I intend to refute these and other arguments in favor of the view that specific purpose language tests are indeed necessary, reliable, valid, and theoretically well-motivated.

Typically, LSP tests have been construed as those involving language for academic purposes and for occupational or professional purposes. Readers may wish to have a look at the following publications for further information on the field of language for specific purposes, of which LSP testing is certainly a part: Swales (1985) for a discussion of the development of the field, and Dudley-Evans and St John (1998) for a discussion of current developments. I will focus on two aspects of LSP testing that may be said to distinguish it from more general purpose language testing: authenticity of task and the interaction between language knowledge and specific purpose content knowledge. Authenticity of task means that the LSP test tasks should share critical features of tasks in the target language use situation of interest to the test takers. The intent of linking the test tasks to non-test tasks in this way is to increase the likelihood that the test taker will carry out the test task in the same way as the task would be carried out in the actual target situation. The interaction between language knowledge and content, or background, knowledge is perhaps the clearest defining feature of LSP testing, for in more general purpose language testing, the factor of background knowledge is usually seen as a confounding variable, contributing to measurement error and to be minimized as much as possible. In LSP testing, on the other hand, as you will see in Chapter 2, background knowledge is a necessary, integral part of the concept of specific purpose language ability.

LSP testing, like LSP teaching, has a relatively short history. A case could be made for the beginning of LSP testing as early as 1913, with the establishment of the University of Cambridge Local Examinations Syndicate’s (UCLES) Certificate of Proficiency in English, a test designed for prospective English teachers to demonstrate their proficiency in the language (University of Cambridge Local Examinations Syndicate 1995). Another candidate for the title of first LSP test might be the College Entrance Examination Board’s English Competence examina-
tion in the US, a test for international applicants to US colleges and universities introduced in 1930 (Spolsky 1995). Both of these tests have clearly defined purposes related to vocational and academic English, respectively, and thus in a sense qualify as examples of LSP tests. However, as I mentioned above, LSP testing requires first, an analysis of a target language use situation, from which characteristics of test content and method are derived, as well as an interaction between language knowledge and specific purpose content knowledge. Clearly, not all examples of what we call LSP tests manage to meet these criteria completely, but I will argue in this book that a theory of LSP testing establishes these two characteristics as fundamental goals. The UCLES and the College Board tests were not developed on the basis of analyses of language teaching or academic situations, nor did the tasks on the tests bear much relationship to the kinds of tasks required of either teachers or students (except when taking language tests!).

So, when might we say that true LSP testing began? A strong candidate is the Temporary Registration Assessment Board (TRAB) examination, a test introduced in 1975 by the British General Medical Council for the purpose of evaluating the professional and language abilities of physicians trained outside the UK applying for temporary registration to practice medicine in Britain (Rea-Dickens 1987). The examination consisted of an assessment of both professional competence and ability to communicate in English. The language component comprised a taped listening test, a written essay, and an oral interview in which both professional knowledge and language ability were assessed. The TRAB language component was based on an analysis of the language, both spoken and written, actually used by physicians, nurses, and patients in British hospitals. As I have discussed, this analytical approach is a critical feature of LSP test development. In addition, the language testing specialists who developed the language component of the TRAB test were not solely responsible for its development, but worked together with medical experts in constructing the tests. This is an important aspect of specific purpose test development. As Rea-Dickins (1987) put it in discussing the TRAB development process, collaboration with practitioners in the specialist area ‘would seem to be a pre-requisite for the design of a “special purposes” test as the domains incorporated within the specialist area go beyond those in which the linguist – independently – is competent to make judgements’ (p. 196). Thirdly, the TRAB developers attempted
to promote the engagement of the test takers’ language ability and background knowledge in the test tasks by providing appropriate and rich contextual features in the test material. For example, in the writing tasks, the test takers were presented with authentic information about a patient’s case history, and the tasks were linked systematically to the problems presented. Typical writing tasks included the following:

Write a letter to Dr Jones summarising the case and giving your recommendations for Mr Brown’s after-care.

Complete the x-ray request card for this examination.

When the patient is admitted to hospital, what written instructions would you leave the night nurse in charge of the ward regarding management?


We can see in this early example of an LSP test the embodiment of the critical features of LSP test development: analysis of the target language use situation, authenticity of task, and interaction between language and content knowledge. The TRAB was later revised (its name changed to PLAB – Professional and Linguistic Assessment Board), and is at present no longer in use, but it stands as a worthy prototype of the art of LSP test development. (Readers might also want to note another early LSP test, the English Language Teaching Development Unit [ELTDU] test, introduced in 1976 as an assessment of vocational English. See North 1994 for information.)

You might reasonably ask the question, however, as to why LSP testing is necessary, or even desirable. To consider this issue, let us imagine a typical language testing situation. As in all good language testing projects, LSP test development begins with a problem to be solved.

A problem

Suppose we want to determine whether people involved in international trade know English well enough to conduct their business. In such a situation, we might reasonably decide to devise a test of English for international business purposes. We would begin our task as test developers by interviewing experienced business people, as well as company supervisors, heads of international divisions, and an
assortment of middle level managers who typically deal with international colleagues. We might observe actual negotiating sessions and business meetings, and tape record participants’ use of English in the various situations they find themselves in: large meetings, one-on-one discussions in offices, individual and conference telephone calls, the ubiquitous business lunch and other business-related social occasions, and so on. Our goal would be to describe the situations in which international business people conduct their work, and the characteristics of the language they use and of the tasks they must perform in English.

We would need to make some decisions about the scope and content of our test. For example, how important is it to test ability to communicate about food or travel? Should we require the test candidates to demonstrate knowledge of their field of business as well as their abilities in English? Such decisions would have to be made in consultation with the sponsors of the test, for their purposes in wishing to give the test – and their willingness to pay for a longer and more varied test! – will help determine what aspects of the milieu of international business we will include in our test. Eventually, however, we would be in a position to produce test specifications, a blueprint of the test we intend to develop, including a statement of the purpose of the test, a description of what it is we intend to measure, a description of the contexts and tasks we intend to include in the test (based on our analysis of the features of the international business domain), details of how the test will be scored, and an indication of how scores on the test should be interpreted.

On the basis of these specifications, we would then actually produce test tasks and assemble a specific purpose test of English for international business. After trying the new test out, perhaps by giving it to a group of business people, and revising it, we would offer it to our target group of prospective international traders. We would interpret their performance on our test as evidence that they could, or could not, use English well enough to succeed in the tasks required of them in the marketplace.

Why bother?

But why go to all the trouble of devising a new test? Why spend the time, effort, and money to interview people, describe the language
tasks of international business, devise the test, and pilot and revise it? Why not just turn to an existing test of English language ability, one such as the Educational Testing Service’s *Test of English as a Foreign Language* (TOEFL), or the Cambridge University Local Examinations Syndicate’s *Certificate of Proficiency in English* (CPE)? These, after all, are well-known international tests, with well-known measurement properties. The TOEFL is a multiple-choice test of listening, structure, reading comprehension, and writing, and is often taken by people who wish to demonstrate English language ability for international communication. TOEFL candidates can opt to take a speaking test as well, to further demonstrate their ability to use English. The CPE is a general test of English reading, writing, structure, listening, and speaking, and is used by many businesses to certify the English language skills of their employees, in addition to its main purpose for university admissions. So, why not use an existing, general purpose language test for our international business candidates?

**Reason 1: language performances vary with context**

One reason is that researchers are pretty much in agreement that language performances vary with both context and test task, and therefore our interpretations of a test taker’s language ability must vary from performance to performance. For example, if we give test takers a reading test based on a passage about square-rigged sailing ships, followed by one based on a passage about micro-chips in computers, they will probably perform somewhat differently on the two tests, particularly if they are studying computer engineering! However, as you will see, it is not enough merely to give test takers topics relevant to the field they are studying or working in: the material the test is based on must engage test takers in a task in which both language ability and knowledge of the field interact with the test content in a way which is similar to the target language use situation. The test task, in other words, must be **authentic** for it to represent a specific purpose field in any measurable way. I will discuss the nature of authenticity in more detail below, but for now let us agree that LSP testing requires the use of field specific content in tasks which might plausibly be carried out in those fields. Returning to our business English example, it would not be enough, in this view, to provide test takers with listening texts about the work of international commerce,
but rather it would be necessary to provide test tasks that share
similar characteristics with the tasks that international traders actu-
ally perform in their work, both in the processing of information and
in responding to it. Thus we must keep in mind that an important
reason for using specific purpose measures is that if we wish to inter-
pret a person’s test performance as evidence of language ability in a
specific language use situation, we must engage the test taker in tasks
which are authentically representative of that situation.

There is quite a bit of research which suggests that this interaction
between the test taker’s language ability and specific purpose content
knowledge and the test task is a necessary condition in LSP tests. It
has been found, for example, that when test takers have some prior
knowledge of the topic of a reading passage, they have an advantage
in responding to comprehension questions based on that passage.
This suggests that there may be no such animal as a pure language
test. Measures of language ability are always colored by such factors
as background knowledge and test method. It has also been found,
however, that the advantage due to specific purpose content knowl-
dge may be quite negligible unless the passage and tasks are suffi-
ciently specific to engage the test takers in authentic language use. I
will discuss evidence for this claim in some detail in Chapter 2.

Reason 2: specific purpose language is precise

A second reason for preferring LSP tests over more general ones is
that technical language – that used in any academic, professional or
vocational field, including cooking, law, physics, chemistry, air traffic
control, scuba diving, religion, stamp collecting, or language teaching
– has specific characteristics that people who work in the field must
control. What we often refer to as jargon or even gobbledygook has a
specific communicative function within that field, namely precision.
There are lexical, semantic, syntactic, and even phonological charac-
teristics of language peculiar to any field, and these characteristics
allow for people in that field to speak and write more precisely about
aspects of the field that outsiders sometimes find impenetrable. It is
this precision that is a major focus of specific purpose language use
and is a major factor arguing in favor of specific purpose language
tests. A classic example of the need for precise, specific purpose
language comes from the field of law. We frequently deplore what we
call legalese, the arcane lexis, the convoluted syntax, the use of Latin terminology, and the interminable cross-references to previous laws and cases in legal texts. Yet, legal language was purposefully developed and is used dynamically by members of the legal profession to communicate among themselves the precise meaning of the law. A good example can be found on the back of any airline ticket:

Conditions of Contract

1 As used in this contract, ‘ticket’ means this passenger ticket and baggage check, of which these conditions and the notices form part, ‘carriage’ is equivalent to ‘transportation,’ ‘carrier’ means all air carriers that carry or undertake to carry the passenger or his baggage hereunder or perform any other service incidental to such air carriage, ‘WARSAW CONVENTION’ means the Convention for the Unification of Certain Rules Relating to International Carriage by Air signed at Warsaw, 12th October 1929, or that Convention as amended at The Hague, 28th September 1955, whichever may be applicable. Ref. 1293 ATB (REV. 6–89)

This statement, not among the most opaque of legalese, but nevertheless quite recognizable as such, was clearly written not by the airline’s public relations officer for the traveler who bought the ticket, but rather by lawyers for other lawyers, and is a good example of the legal profession’s demand for precision in language. If, for whatever reason, we wanted to measure a lawyer’s control of English to conduct the business of law, it would not seem to be sufficient to use texts and tasks which were not specific to the legal profession. There may be perfectly good reasons to include language and tasks not so strictly related to the legal register in the test, but certainly if our goal is to measure a test taker’s ability to use language within a specific vocation, profession, or academic field, and that is the focus of this book, then specific purpose texts and tasks will be needed.

How are specific purpose language tests related to other types of language tests?

Speaking of precision, it is, of course, necessary to be more precise about the nature of specific purpose language tests than I have been so far. For the moment, let us agree to define our object of interest as tests which attempt to measure language ability for specific
vocational, professional, and academic purposes. I will suggest a more precise definition later in this chapter, but before we can arrive at a useful definition of specific purpose language testing, we need to discuss a number of related concepts in language testing that form the background to LSP testing. These include communicative testing, general proficiency testing, criterion-referenced testing, and the notion of authenticity.

Communicative tests

Particularly since the publication in 1978 of Widdowson’s book, *Teaching language as communication*, and in 1980 of Canale and Swain’s paper, ‘Theoretical bases of communicative approaches to second language teaching and testing,’ the related fields of language pedagogy and language assessment have been characterized by the communicative paradigm, the communicative approach, and communicative language teaching. But even before the publication of Hymes’s (1972) classic paper, ‘On communicative competence,’ which provided much of the impetus for the communicative approach, language testers were discussing ‘productive communication testing’ (Upshur 1971), and teachers and testers have been fascinated with the notion for over a quarter of a century now. As you will see below, specific purpose language tests are by definition communicative. Indeed, Sajavaara (1992), in a discussion of LSP test design, assumes from the outset that ‘It is impossible to distinguish LSP testing theoretically from communicative language testing’ (p. 123).

In his book *Communicative language testing*, Weir defines his topic as follows:

> In testing communicative language ability we are evaluating samples of performance, in certain specific contexts of use, created under particular test constraints, for what they can tell us about a candidate’s communicative capacity or language ability.

Weir (1990: 7)

In his definition, Weir employs a number of key terms: communicative language ability, specific contexts of use, test constraints, and capacity. Since specific purpose language testing involves all these concepts, we will conceive of it as a special case of communicative language testing. The first of Weir’s terms, communicative language
ability (CLA), was introduced by Bachman as a framework for describing language knowledge and the capacity for implementing it ‘in appropriate, contextualized communicative language use’ (Bachman 1990: 84). This leads us to Weir’s second key term, specific contexts of use, requiring us to take account of the many features of context that influence communication, features such as the physical and temporal setting, the role(s) of the test taker and the interlocutor(s)/audience, the purposes of the communication, the topic and content of the message, its tone and manner, and the channels, codes, and genres being employed (cf. Hymes 1974). The third key term in Weir’s definition, test constraints, reminds us that the methods we employ in eliciting a language performance will influence the nature of the performance and thus the interpretations we might make on the basis of it. Tests are, after all, contrived language use events, and even the most cleverly contrived test tasks limit to some degree the generalizability of our interpretations concerning the test takers’ specific purpose language abilities.

Finally, Weir refers to capacity, a term employed by Widdowson (1983), as ‘the ability to use knowledge of language as a resource for the creation of meaning’ (p. 25), and is intended to be understood from the perspective of the language user rather than that of the language analyst (or, indeed, the language tester). In this book, I will use the term communicative language ability (and later, specific purpose language ability) to capture the notion of capacity as Weir and Widdowson use the term. The point that is crucial in the testing of language ability in specific purpose contexts is understanding that ability from the perspective of the language user. That is, not only are we interested in measuring communicative language ability rather than language performances per se, but we are called, in LSP testing, to interpret test performance from the point of view of language users in the specific purpose situation of interest. Thus, specific purpose language testing, as Widdowson points out with regard to specific purpose language teaching, is essentially an exercise rooted in an understanding of human activity from the point of view of the participants in the activity. In this regard, in Chapter 2, I will explore the concepts of grounded ethnography and indigenous assessment, as useful approaches for understanding the nature of LSP test performance from the point of view of the language users.
General purpose tests and specific purpose tests

Defining purpose

Widdowson (1983) points out that although, as I said above, all language courses (and tests) are purposeful, there is a difference in how purpose is defined. He suggests that in general purpose language courses, a distinction is made between aims, the eventual target behaviors of the learners, and objectives, pedagogical constructs which, it is believed, will enable the learners to achieve the behavioral targets. The goal, he says, of general purpose language courses, is to provide learners with an ability to solve on their own the profusion of communication problems they will encounter when they leave the language learning classroom. On the other hand, designers of specific purpose language courses, Widdowson suggests, often collapsed the distinction between aims and objectives, so that descriptions of target behaviors, usually derived from a needs analysis of a specific purpose language situation, become the course content. In other words, Widdowson argues, specific purpose language teaching suffered from a lack of theoretical motivation for course design, and became a very narrowly focussed training exercise in which learners were taught specific behaviors but not strategies enabling them to adapt to new, unspecifiable situations. Although many LSP courses are nowadays more strategically oriented, language teachers can tell anecdote after anecdote about learners who demonstrate an ability to perform the required language functions in the context of the classroom, but as soon as they walk out of the door, fail miserably to transfer the skills to the requirements of communication outside the classroom. There seems often to be a gap between what students can do in the classroom and what they can do in the real world.

In discussing the issue of a lack of theory in LSP testing, Davies (1990) argues that ‘Tests of LSP/ESP are indeed possible, but they are distinguished from one another on non-theoretical terms. Their variation depends on practical and ad hoc distinctions that cannot be substantiated’ (p. 62). It is one of the purposes of this book to provide some theoretical justification and frameworks for LSP testing that will take it out of the realm of narrowly focussed behavioral assessment and bring it more in line with the theoretical underpinnings of communicative language testing. Such an approach will lead to the assessment of the abilities that underlie communicative performance which
will be generalizable from one situation to the next and from the test situation to non-test target situations. For example, if test takers demonstrated in a test that they could successfully read a patient’s case history and then could use that information to write a letter of referral to another physician, we want to be certain that they could carry out a similar task in a non-test situation. More importantly, we must ask how many different communicative situations might the test takers potentially need to deal with in their work as physicians which need to be tested in a test of medical English. However, it is practically impossible not only to sample the variety adequately, but even to list all the possible permutations of communicative events that language users must cope with. The problem of generalizing to real life is a central one in LSP testing. We wish to avoid the potential problem of producing a test on which performance is interpretable only in terms of that test. This situation can come about when we equate target behaviors with test content without noting the features of the target situation that are shared by the test tasks. As I will discuss below, it is only by taking note of the features of the target situations and comparing them with those of the test task, that we can make that inference with any certainty. In this discussion, and, indeed, in the rest of the book, I will draw heavily on Bachman and Palmer’s book *Language testing in practice* (1996), for their approach to test design and development is one that is useful in informing and carrying out any LSP testing enterprise.

*Generalizing to real life*

It has proven very difficult, and may eventually prove to be impossible, to make predictions about non-test performance in the real life target situation on the basis of a single test performance, no matter how true to real-life the test tasks might be. This is so because language use, even in highly restricted domains, such as taxi-driving, accounting, welding, biochemistry, or waiting tables, is so complex and unpredictable that coverage, or sampling of tasks, will be inadequate. Skehan (1984), for example, writing about the problems of testing English for specific purposes, notes that ‘Merely making an interaction “authentic” does not guarantee that the sampling of language involved will be sufficient, or the basis for wide ranging and powerful predictions of language behaviour in other situations’ (p. 208).
Spolsky (1986) agrees, and reminds us that how speech acts are realized is the result of a complex interaction among many contextual variables, and although we might study pragmatic values and sociolinguistic probabilities of various forms appearing in different contextual environments, ‘the complexity is such that we cannot expect ever to come up with anything like a complete list from which sampling is possible’ (p. 150). Bachman (1990) points out that now that it has become commonplace to recognize that language use takes place in contexts, and must be interpreted with reference to the context, and since the domain of language use consists of a potentially infinite number of unique instances, the assumption that we will be able to predict future communicative performances on the basis of a single test performance becomes untenable. He offers an example of attempting to produce a test of English proficiency for taxi-drivers in Bangkok by making lists of actual utterances the drivers might be expected to control. It soon became clear that the complexity involved in negotiating meaning even in this relatively narrowly defined context meant that ‘there was probably an infinite variety of conversational exchanges that might take place’ (p. 312). Skehan hypothesizes a similar problem in another domain, that of a waiter in a restaurant:

Although at first sight ‘waiter behaviour’ might seem to be a straightforward affair, we soon need to ask questions like: what range of customers needs to be dealt with? What range of food is to be served? Once one probes a little, the well-defined and restricted language associated with any role is revealed to be variable, and requiring a range of language skills.

Skehan (1984: 216)

Tests developed in the real-life mold, which equate language ability with a specific language performance, are analogous to the training courses criticized by Widdowson, above, as failing to test the ability of the learners to deal with new, unexpected, unique communication problems. This is a real problem for specific purpose language testing. Tests might contain tasks that mirror faithfully those of the target situation, and these tasks might meaningfully engage the test takers’ language ability, and yet the test overall might not be truly representative of the target situation, since there are simply too many possible variations of target situation to cover adequately in a test.

I will consider this problem in more detail in the discussion of authenticity later in this chapter, but for the moment let us agree that
what is required in LSP testing is not the holistic replication of a specific purpose domain, but rather the use of features or characteristics of tasks in specific purpose language use situations in the construction of test tasks. This leads us to a view of LSP testing in which test tasks are developed on the basis of an analysis of characteristics of context and tasks in target language use situations. It is this analysis of target language use task characteristics which will allow us to make inferences about language ability in the specific purpose domain. The distinction between ability and performance is an essential one in the approach to language testing advocated in this book. The interaction between ability and task characteristics leads to authenticity, which I will interpret as the extent to which the test does in fact engage the test takers in tasks characteristic of the target language use situation. It should be clear from this discussion, too, that language tests are not either specific purpose or general; rather, there are degrees of specificity, which can be described along two dimensions: the amount of content or background knowledge required for carrying out test tasks, and the narrowness of interpretations which may be made on the basis of test performance about language use in real-life contexts. In other words, language tests will be more or less specific purpose in relation to the degree to which they require the engagement of specific purpose content knowledge in responding to the test tasks and the degree to which they allow generalizations about language use in specific situations. For example, the Test of English as a Foreign Language (TOEFL) is a test intended to measure English proficiency broadly interpreted, without the engagement of any special background knowledge or specific reference to use (Educational Testing Service 1965), and would thus be considered a more general purpose language test. On the other hand, a test such as the Proficiency Test in English Language for Air Traffic Controllers (PELA), a language test for trainee air traffic controllers in Europe (Institute or Air Navigation Services 1994), requires a large amount of specialized knowledge about air traffic control, and interpretations of language use are specifically limited to the work of air traffic control officers. The PELA is therefore a prototypical example of a highly specific purpose language test. Between these two extremes is a test such as the Test of English for Educational Purposes (TEEEP), intended as a test of academic English skills (Associated Examining Board 1984), though not related to any specific field, and the Taped Evaluation of Assistants' Classroom Handling (TEACH), a
test of instructors’ ability to present information to students in specific academic fields (Abraham and Plakans 1988). I will return to discussion of the characteristics of target language use situations and language test tasks in Chapter 5 and will consider these and other tests in Chapters 6 and 7 in light of such characteristics. For now, I will merely emphasize that test developers must always take both test purpose and task characteristics into account when setting out to measure communicative language ability.

**Criterion-referenced tests and norm-referenced tests**

A very important concept for specific purpose language testing is that of criterion-referenced (CR) testing. Usually contrasted with norm-referenced (NR) language testing, CR testing differs from NR tests both in design and in the interpretation we make of performance on them. NR tests are designed to maximize distinctions among test takers so as to rank them with respect to the ability being tested; CR tests, on the other hand, are designed to represent levels of ability or domains of content, and performance on them is interpreted with reference to the criterion level (Bachman 1990). In other words, on NR tests, passing would be determined by relative ranking within the population of test takers, while on CR tests, test takers might all achieve the criterion and so pass. As an example of CR test use, suppose we wanted to test prospective candidates for certification as scuba divers to see whether they knew certain essential diving terminology, such as bc jacket, pony bottle, and regulator. Scuba divers use this terminology in pre-dive checks, so it is essential to know whether candidates can carry out a pre-dive check using the appropriate vocabulary. We are not interested in discovering who knows the most terms; we want to know which candidates know all the essential terms. Our performance criterion, then, is whether the candidate can use all the terms appropriately. On the other hand, continuing with the scuba diving example, as a way of motivating students to learn the various concepts associated with diving, such as the relationship between depth and pressure, and the different types of equipment, an instructor might offer a weekly prize to the five students who get the highest scores on quizzes. In this case, the aim would be to rank the students against each other so that the top five could be identified – a norm-referenced use of the tests. Both types of test uses are relevant
to LSP testing; however, the development process associated with CR
testing, which involves a detailed analysis of the target language use
situation, is of most direct relevance in LSP testing, particularly with
regard to a fundamental concept in specific purpose testing, authentici-
city.

It is important, therefore, to note that specific purpose language
tests might be developed as either CR or NR tests, but CR testing
offers an important perspective to LSP testing: the necessity of speci-
fying precisely the level of ability or the domain of content that is to
be the criterion for performance. Thus, the process of developing a
CR test, requiring as it does the precise, detailed specification of not
only the features of the specific purpose target language use situation,
but also the criteria for evaluating performance, is extremely useful in
LSP test development. In fact, the development of evaluation criteria,
or rating scales, is perhaps the most important, and also the most
Until very recently, the task of developing assessment scales has been
left to test developers and other applied linguists, and, not surpris-
ingly, the scales they have come up with reflect a linguistic orienta-
tion, so that such categories as grammar, cohesion, vocabulary,
fluency, intelligibility or comprehension are commonly employed. I
will discuss problems with this approach in Chapter 5, and suggest a
possible solution, but for now, let us simply note that a precise defi-
nition of assessment criteria is an essential part of the LSP test develop-
ment process, and CRT procedures offer a systematic approach to
specifying these criteria.

**Authenticity**

Since authenticity is such an important concept in specific purpose
language testing, it is necessary to consider its meaning in some
detail and with some precision. Kramsch (1993) points out that the
term has been used to indicate a reaction against the often artificial
language of language textbooks and tests; it refers to the way
language is used in non-pedagogic, non-test, natural communication.
Since the publication of Widdowson’s *Explorations in applied linguis-
tics* (1979), many language teachers and testers have come to view
authenticity as a property not of spoken and written texts themselves,
but of the uses people put them to:
It is probably better to consider authenticity not as a quality residing in instances of language but as a quality which is bestowed upon them, created by the response of the receiver. Authenticity in this view is a function of the interaction between the reader/hearer and the text which incorporates the intentions of the writer/speaker . . . Authenticity has to do with appropriate response.

Widdowson (1979: 166)

For example: a set of instructions for conducting a chemistry laboratory exercise may be a perfectly authentic piece of material, but when used in a multiple-choice language test as a vehicle for testing knowledge of vocabulary or the use of imperatives, it is not being used for the purpose intended by the author of the chemistry lab manual, or in the way lab supervisors would use it. A key concept in Widdowson’s formulation above is that of interaction between the language user and the text, and I will make use of this notion in my characterization of authenticity in specific purpose language testing.

In response to this problem for language teachers, Widdowson (1979, 1983) notes that there is often a confusion between the use of ‘authentic’ to refer to examples of language actually produced by users in a communicative situation versus reference to the activities and procedures that language users engage in, in association with the forms of language produced. He suggests a distinction between the terms authentic and genuine: the former referring to activities or processes associated with instances of language use, and the latter for the actual spoken or written texts produced by the users. Thus, our use of a set of instructions from a lab manual for the purpose of testing an instructor’s ability to understand and use imperatives would be the use of a genuine text for a purpose other than that for which it was intended. Bachman (1991) reminds us of Widdowson’s point, quoted above, that authenticity is a function of an interaction between a language user and a discourse, and proposes two aspects of authenticity: situational and interactional. The first aspect is composed of authentic characteristics derived from an analysis of tasks in the target language use situation, the features of which are realized as test task characteristics. Thus, situational authenticity can be demonstrated by making the relationship between the test task characteristics and the features of tasks in the target language use situation explicit. The second aspect of authenticity, interactional, is closely related to Widdowson’s definition above, and involves the interaction
of the test taker’s specific purpose language ability with the test task. The extent to which the test taker is engaged in the task, by responding to the features of the target language use situation embodied in the test task characteristics, is a measure of interactional authenticity. It is important in specific purpose language tests that both these aspects of authenticity are present. It is quite possible, for example, that a test task may be perceived by test takers as having nothing whatever to do with their field of study, but which they nevertheless find quite interesting and which engages their communicative language ability interactively. Performance on the task would be interpretable as evidence of their communicative language ability, but not in the context of the target language use situation. By the same token, a test task may contain all the contextual attributes of the target situation and yet fail to engage the test taker meaningfully in communicative language use. Mere emulation of a target situation in the test is not sufficient to guarantee communicative language use, and, as Lewkowicz (1997) has pointed out, the focus on the interaction between the test taker’s language ability and the situational characteristics of the test task is a strength of this dichotomous view of authenticity in specific purpose language tests.

I will develop these concepts more fully in Chapter 3, but for now, I propose to employ this dual notion of authenticity in specific purpose language testing. In LSP test development, what we must do is first describe a target language use situation in terms of features of context and task; we must then specify how these characteristics will be realized in the test so as to engage the test taker in test tasks, performance on which can be interpreted as evidence of communicative language ability with reference to the target situation. Building on the work of Bachman and Palmer (1996), Bachman et al. (1991), and Davidson and Lynch (1993), I will develop a ‘means of classifying test tasks on the basis of dimensions . . . that we abstract from authentic language use’ (Bachman 1990: 317) in the construction of specific purpose language tests.

**Specific purpose language tests**

I am at last ready to propose a more precise definition of specific purpose language tests than that I suggested at the beginning of the discussion.
A specific purpose language test is one in which test content and methods are derived from an analysis of a specific purpose target language use situation, so that test tasks and content are authentically representative of tasks in the target situation, allowing for an interaction between the test taker’s language ability and specific purpose content knowledge, on the one hand, and the test tasks on the other. Such a test allows us to make inferences about a test taker’s capacity to use language in the specific purpose domain.

This definition has emerged from the discussion of a number of concepts in this chapter. I have discussed reasons for wishing to develop specific purpose language tests, and noted first that language performance varies with both context and test task, and that therefore our interpretations of a test taker’s language ability must vary from situation to situation. Second, we have seen how technical language – that used in any academic, professional or technical field, including cooking, law, physics, chemistry, air traffic control, scuba diving, religion, or stamp collecting – has specific purpose characteristics that people who work in the field must control. What we commonly refer to, often disparagingly, as jargon has a specific communicative function within that field, namely precision.

In considering how specific purpose language testing is related to other types of and approaches to language testing, I discussed the distinction between so-called general tests and LSP tests and we saw that while all tests have purposes, in LSP testing, the notion of purpose is typically more narrowly focussed than in more general language testing. There is a problem inherent in this focus, however, since there is in principle no way of determining how specific specific needs to be. A criticism of specific purpose testing has been an assumption that if a test taker could perform the real-life test task, he or she would be able to perform in the target language use situation. However, there are serious problems in demonstrating this to be the case. It is impossible, except in the most restricted language use situations, to specify with any completeness the range of language forms that will be required. This is so because language use, even in relatively specific domains, is so complex and unpredictable that coverage, or sampling of tasks, will be inadequate.

As a way out of the dilemma of never-ending specificity on the one hand and non-generalizability on the other, I have referred to context and task characteristics, which are drawn from an analysis of a target language use situation, and which will allow us to make inferences
about language ability in the specific purpose domain. In specific purpose language test development, what we must do is first describe a target language use situation in terms of characteristics of context and task, then specify how these characteristics will be realized in the test so as to engage the test taker in test tasks, performance on which can be interpreted as evidence of language ability with reference to the target situation.

I should discuss briefly here a concept that I have used a number of times in this chapter but have not yet defined precisely: inference. A central goal in language testing is making judgements about test takers on the basis of their performance on a test. That is, we give tests to elicit performances which we can observe so that we can make inferences about qualities of test takers which we cannot observe. A fundamental question involves what we wish to make inferences about. We might want to make inferences just about a person’s language ability; for example, this candidate is able to write business letters in English, using correct syntax, vocabulary, and spelling. Alternatively, we might want to make a statement not only about language ability, but also, particularly in LSP testing, about specific purpose background knowledge; for example, this candidate is able to write business letters in English, incorporating appropriate types and amounts of information from material provided. Inferences of this type would be more complicated, since we would need not only a measure of language ability, but also one of background knowledge, so as to be able to disentangle the two types of knowledge and understand, for example, whether a candidate’s failure to incorporate appropriate types and amounts of information was due to a lack of language ability or a lack of background knowledge. Finally, we might wish to make inferences about a candidate’s specific purpose language ability, in which case, language and background knowledge would be left intertwined. The type of inference we want to make would depend on the purpose for which we were giving the test, but in specific purpose language testing, the first type of inference, that about a decontextualized language knowledge, is probably not very useful. The second type, where we want to separate out language knowledge from background knowledge, would be most useful when, for example, the test takers were trainees in the specific purpose field, and we needed to know if their low test performance was the result of problems with the language or a lack of background knowledge, so we could offer appropriate remediation. The final possibility, inferences
about the dual component specific purpose language ability, would be most useful in situations where we could take the test takers’ specific purpose background knowledge for granted, as in the case of qualified doctors who wish to demonstrate their language ability for purposes of licensure.

Making appropriate inferences is a crucial aspect of specific purpose language testing, and I want to discuss it just a bit more deeply. McNamara (1989, 1996) has distinguished between making inferences on the basis of LSP test performance about ability to do future tasks or jobs in the target language use situation, on the one hand, versus making inferences about ability to use language in specific future tasks or jobs, on the other. This seems a subtle distinction, but it is of extreme importance for the theoretical foundations of LSP testing. McNamara cautions against the first type of inference since job performance is influenced by a number of factors, such as personality characteristics, that are independent of language ability. He makes a theoretical distinction between a strong performance hypothesis, about an individual’s ability to perform target tasks successfully, and a weak performance hypothesis, about ability to use language in the target situation, and prefers the latter. I will argue in Chapter 2 that we are not attempting to measure communicative success in LSP tests, but rather the knowledge and abilities that underlie communicative performances, and this point is related to McNamara’s. I have asserted that it is practically impossible not only to sample the variety of tasks in a target domain adequately, but even to list all the possible types of communicative events that language users must cope with. Thus, I agree with McNamara that we should restrict ourselves in LSP testing to making inferences about language ability and not about job performance. However, I have also tried to establish a case for making inferences about specific purpose language ability, a construct defined on the basis of an interaction between language knowledge and specific purpose background knowledge. This departs somewhat from McNamara, who would include job-related background knowledge in his list of factors essentially unrelated to language knowledge. I will also argue later that language knowledge must be interpreted differently from one domain of use to another, since as Chapelle (in press) points out, context constrains language choice, and if we are interested in making inferences about test takers’ abilities to use language in specific situations, then background knowledge associated with