1 Introduction

This book is based on a study centring on a test of speaking. However, the test itself – a 20-minute communicative test, conducted in pairs, for lower secondary school Norwegian students of English – is not really what the book is about. It is a fairly unremarkable test, of the type that anyone conversant with generally accepted principles and practice concerning the testing of spoken interaction, evolving around the turn of the millennium, might have produced, given resources and a relatively free hand. What is, I trust, of interest to the reader is what emerges from the book on the actual validation of the test and on a particular body of language – 'smallwords' – which actually seems to characterise the speech of more fluent speakers of English. This dual focus takes the study beyond the particular test and provides the reader with frameworks both for the testing of any test of spoken interaction, and for investigating the fluency/smallword use of any learners s/he may be concerned with.

Test validation

Validating a test really means attempting to answer the simple question 'does it work the way it is intended to?' The value of being able to answer this is obvious, whether we are looking at a test that is already in operation, or whether we are embarking on designing or choosing a test for future use. The answer, of course, is rarely simple, and finding it is even less so! A test involves many processes, from the original decision to have a test, through all the stages in making it and carrying it out, to the uses that are made of the results. And things can happen at any point that may send it off course. How, then, can we keep track of a test, checking it for damage as it moves through this minefield?

The literature on test validation is vast, and the number of 'types' of validation addressed has increased dramatically in the last decade or so; four classical types are cited in Hughes (1989), while Cummings (1996) lists 16. At the same time, there is a move towards accepting only one, unified, validity, championed by Messick (e.g. 1996). Although much of the discussion is invaluable in giving the reader theoretical, and often practical (e.g. Alderson *et al*. 1995), insight into what makes a test work, it is difficult to find any clear, systematic way of actually testing our tests from start to finish.

The first part of the study detailed on in this book attempts to provide a framework for doing just this, and demonstrates it in use. By combing through the literature for a consensus on what seems to threaten validity in language testing, it attempts to isolate all the factors that can make a test go wrong, and to further identify these as threats to any of six basic aspects of validity, building on Messick (1996). It goes on to apply this framework to the test in question, examining the test itself, 'as it stands', as well as the data emerging from the test in use. The result is a comprehensive estimate of the state of the test, showing what seems to be functioning satisfactorily, what needs further investigation and what seems to be malfunctioning.

What comes out of this estimate provides the background for the second part of the book. A major flaw in the test was found to lie in its band-scale descriptors of performance across levels, along with the profile form, which is the basis for setting the level on the scales, particularly those parts associated with what is conventionally termed 'fluency', as opposed to 'language'. Here, among other things, too little reference was made to linguistic markers of fluency (i.e. actual items of language). Moreover, this neglect seemed to go hand in hand with the virtual absence of reference to *smallwords* (small words and phrases that contribute to the act of speaking rather than to the message itself, such as *you know, well, right*). As the primary purpose of the testing is to provide detailed, pedagogical feedback of learners' strengths and weaknesses, through the band scales and the profile form, these shortcomings had to be taken seriously. The remainder of the study attempts to redress the flaw, by focusing on fluency and on the potential role of smallwords in bringing this about.

Fluency and smallword use

The first task was to establish an explicit theoretical link between fluency and the use of smallwords, by reference to the literature on both of these themes. Next, using an electronic corpus of the transcripts of test takers (both Norwegian and native-speaker students), the speech of students at different levels of fluency (based on grade or native-speaker status and backed up by data on temporal markers) was contrasted for smallword use. In terms of quantity and range, there was found to be no doubt that fluency and smallword use were correlated. However, it was necessary to investigate the actual way smallwords were used. As in the case of validation, the literature was only partially helpful, and a framework had to be devised for analysing smallword use. *Relevance theory* (Sperber and Wilson 1995) was drawn on, giving rise to a five-macrosignal framework, within which each smallword was able to be classified for the signal(s) it was sending, the data being contrasted across the groups. Clear tendencies were found in the students' language, showing

a gradual acquisition, as fluency increases, of native-speakerlike signals sent by smallwords.

The findings provide a basis for the writing of new, data-driven descriptors of 'fluency', with explicit reference to smallword use, of the type called for by Fulcher (1996). Not only does this give the potential to remedy a flaw in the test under scrutiny, but it also contributes to the pool of corpus-based knowledge of what goes on in the speech of younger learners at different fluency levels, measured against the yardstick of the speech of native speakers of the same age.

The test

The test that is validated here, and which is the source of all the data analysed, is the speaking test part of the EVA (Evaluation of English as a school subject) diagnostic test material for 14–15 year olds in Norwegian schools. This material, sponsored by the Norwegian Ministry of Education, was developed in the University of Bergen English Department, and piloted nationally in the spring of 1995, prior to going into full operational use. The primary purpose of the testing was defined as providing teachers and students with detailed information on the strengths and weaknesses in students' communicative language ability (CLA) in English, so that learning activities could better be adapted to students' particular needs. The methods used were to be innovative and the process of taking part in the testing was intended to enhance the learning situation and the level of competence in assessment itself.

The speaking test consists of a set of tasks and scoring instruments. The tasks (Appendix A) are carried out by students in pairs, and involve describing, narrating, instructing and semi-role-play. There are three parallel versions of the test. There are two types of scoring instrument: a performance profile form (Appendix B) and a pair of band scales (Appendix C). The profile form contains a number of detailed, closed-answer questions covering many aspects of performance. When completed, this should be used as a guideline for setting a level on each of the two band scales *message and fluency* and *language structures and vocabulary*. Raters must first place the student at one of six levels on both scales, and then award a global grade on the basis of these placings, taking pronunciation and intonation into account as a final adjuster (Appendix D).

Research questions

The research questions, which are summed up in this section, can be divided according to whether they are addressed at a theoretical or an empirical level. Theoretical issues concern firstly test validation, reviewing recognised causes

and effects of invalidity. The discussion then arises of what makes up CLA in the case of the students being tested. Later, the question of 'what fluency is' is opened up, both in terms of the 'surface effects' of fluency and its underlying 'causes' or the skills necessary to bring it about. The focus then moves to smallwords themselves, seeking to identify the signals they send and to show a correspondence between these signals and the skills underlying fluent speaking. A significant aim of the study is to provide a framework within which any smallword (as defined here) may be analysed in terms of the signals it sends. Finally the learner is put under the spotlight, posing questions of how smallwords might be acquired and fluency strengthened.

The empirical questions also relate to both the validation of the test and the link between smallwords and fluency. Seven principal empirical questions are posed, which form the backbone for the analyses in the research:

- Which aspects of validity appear to be at risk in the test 'as it stands', and what are the likely causes of invalidity?
- How far do raters' scores provide actual evidence of this suspected invalidity, and shed further light on its causes?
- Is there evidence in the corpora of non-linguistic, temporal features (such as filled pausing) which lends support to the grouping of students into more and less fluent speakers on the basis of their test grades?
- Is there evidence in the corpora that the more fluent learner group used smallwords quantitatively in a more nativelike way than the less fluent group?
- Is there evidence in the corpora that the more fluent learner group used smallwords qualitatively in a more nativelike way than the less fluent group?
- How might these findings be applied to the assessment of fluency?
- Can these findings ultimately be applied to raise the level of fluency in learners?

Data and methods

In order for the reader to understand roughly how the research questions are addressed, it is necessary to give a brief overview of the data and methods used in the research.

The data can be regarded as being made up of three parts. Firstly, there is the test itself, consisting basically of tasks and scoring instruments (see Appendices A to D). Secondly, there is scoring data from a group of raters (two per student) for 59 students. This data consists principally of the global grades, based on the joint 'level' on the band scales, and the scores on the individual questions on the performance profile. Certain other data, e.g. biodata and teachers' and students' estimates of ability, are also available

for most of the students. Thirdly, there is a corpus of the transcripts of the test performances of these same students, as well as a control corpus of transcripts of native-speaker students doing the same tasks. The corpora are accessed through the Internet, using a TACTweb search programme. The tapes and transcripts in hard copy are also available. This data has all been collected (apart from the control corpus) from the national piloting round of testing, which means that the data is based on 'genuine' testing.

Several methodological approaches are employed in the research. On one hand I have done a considerable amount of delving into literature, in order, for example, to work out definitions of what goes into test validation, what makes up CLA, how fluency is described and which signals may be assigned to smallwords.

On the other hand, I have carried out a good deal of data analysis. The more quantitative of these analyses include those using scores, e.g. to test inter-rater reliability, and those using corpus data to make cross-group comparisons of the frequencies of occurrences of features such as smallwords. In these analyses, statistical testing of varying kinds (and with varying degrees of confidence) is employed. Factor analysis is also used, to explore the way aspects of performance cluster. Statistical tests are interpreted as giving indications rather than certainties. Other analyses are more qualitative, frequently involving fitting items into a theoretical framework, such as that of CLA, or that of the signals that can be sent by smallwords.

No absolute claim is made to extrapolate the findings from this dataset to fit all learners (or even all performances of these learners). However, the relatively large size of the datasets and the randomness of the selection of students taking the test, as well as the fact that the initial expectations tend to be corroborated, lend encouragement to the belief that what is found here is probably representative of teenage learners in the Norwegian context. The reader will, of course, form his/her own conclusions on both the validation process and the fluency/smallword use study. The findings, however, are concrete and the methods are largely replicable; moreover, two frameworks are provided, the first for use in test validation and the second for analysing smallword use. What is presented here can thus be regarded as a starting point for others to use as it is, or to build on further or to adjust, whether their primary interest lies in testing or in learner language, or both.

Organisation of the book

Within the two main parts – corresponding to the two central themes of test validation and fluency and smallword use – this book is divided into ten chapters. Following this introductory chapter, Part One 'Test validation' consists of the next four chapters, which cover the processes of validation

applied to the test in hand. In Chapter 2, the notion of validity, as it has been presented in the literature of language testing, is reviewed and a framework for systematic validation is evolved. In Chapter 3, the validation begins by taking up the questions of what constitutes CLA, and how this is defined and put into operation. In Chapter 4, a systematic search is conducted for potential causes and effects of invalidity in the test 'as it stands', yielding a preliminary profile of the test's validity. Chapter 5 investigates the extent to which scoring (and other) data bear out conclusions reached in Chapter 4. Conclusions are drawn as to which principal sources of invalidity exist and how these affect the test, and as to what are the most critical needs to be addressed, now and in the future, in order to enhance the test's validity.

Part Two 'Fluency and smallword use' includes the remainder of the body of the book, i.e. Chapters 6 to 9. This part aims firstly to establish a link between fluency and smallwords, and secondly to examine student transcripts for empirical evidence of both non-linguistic and linguistic (i.e. smallword) markers of fluency. Chapter 6 begins with a discussion, based on recent literature, of what is meant by 'fluency', culminating with the proposal that fluency in speaking is marked by both temporal and linguistic features, the latter being notably the use of smallwords. The second part of the chapter takes a deeper look at fluency and the role of smallwords in promoting this, working towards the building up of a *relevance-theory* framework of the signals sent by smallwords within which their use can be analysed. Chapter 7 uses corpus linguistics, firstly to establish whether recognised non-linguistic markers of fluency are found to support the grouping into more and less fluent performances, judged by the global grades allocated by raters. Secondly, a comparison is made of the extent and distribution of smallwords in the performances of more and less fluent learners and native-speaker students. Chapter 8 employs the framework worked out in Chapter 6 to analyse the way the various student groups assigned signals to smallwords in their speech. Chapter 9 puts the smallword user into focus, and attempts to address the salient issues of what might cause individual variation in smallword use, how smallwords might be acquired, and what the implications of the study are for the teaching and assessment of foreign languages. The book culminates in a conclusion in Chapter 10.

In order to clarify the way terminology and abbreviations are used, relating either to language testing or to analysis, a glossary is provided at the end of the book.

Part One: Test validation

2 Test validation

This book is (largely) about achieving valid testing. Whether or not a test is valid hinges on the question 'does the test test what it is supposed to test?', cited by Alderson *et al.* as 'the most important question in all language testing' (1995: 170). And the reason why this question is so important is that, even though many things may cause them to be flawed, tests are taken seriously, and their results are usually believed and acted upon in some way.

This puts a burden of responsibility on test makers, who are forced to be explicit by the nature of testing, while handling two concepts – language ability and measurement – that are rife with uncertainty. As Davies (1990) puts it:

> language testing compels *explicitness* about language, about language learning, language teaching language performance. [...] It requires us to spell out in detail language criteria, language needs and language levels – not merely so that we can judge whether they have been met or reached but also so that we can explain to others what they mean. Language testing operationalises subjective judgements and in doing so both clarifies and validates them. But the explicitness of language testing – we have called it its main value – exacts a price, the price of *uncertainty*. Language tests do not provide exact information, it is always 'more' or 'less' and 'within confidence limits'. (1990: 53)

While it may never be possible to be certain that a test is testing what it is supposed to, we are able to take steps to reduce our uncertainty. That is what validation is about. This chapter looks into the question of what may cause language tests to be flawed, so that any serious sources of invalidity in the current test can be tracked down and ultimately put right. Exposure to a long-term process of validation will enable the test to be used with an increasing amount of confidence that what can be inferred from its results is more or less true.

This chapter consists largely of an overview of what is generally regarded as comprising validation. Different types of validation are described, and specific sources of invalidity are identified. The overview culminates in presenting a unified approach to validation, whereby potential sources of invalidity are summed up and placed in a theory-based framework. This enables us to see not only which factors pose a threat to validity, but also how they combine to affect validity in certain distinguishable ways. The chapter concludes by outlining how this framework is used to structure and guide the validation process in the remainder of the first part of the book.

Validation – an overview

Hughes' (1989) statement: 'a test is valid if it measures accurately what it is intended to measure' (1989: 22) seems to capture the essence of validity as it has been described in the testing literature. However, as has been emphasised, e.g. by Henning (1987), Bachman (1990), Messick (1995) and many others, there is no such things as a valid test *per se*, because validity is always relative to the purpose of the test: a test may be valid for one purpose, but not another. And Bachman couples this point with a further one: 'To refer to a test or test score as valid, without reference to the specific ability or abilities the test is designed to measure *and* the use for which the test is intended, is [...] more than a terminological inaccuracy' (1990: 238). In other words, the process of validation must begin by establishing what it actually is that is being tested and why the test is being given (and how it will be used).

The 'thing' being tested in a language test is some sort of language ability, used in some domain. It may be a restricted part of ability, e.g. grammatical, or used in a restricted domain, e.g. business language. The ability and domain need to be defined at an abstract level, either by referring to syllabuses and course material, or by drawing on a theory-based description of language ability, or a combination of both.

Any test that claims to measure 'ability' must be founded on an underlying theoretical model, or 'construct', consisting of components of ability. To make this model usable, these components have to be operationalised in terms of actual language behaviour which may be regarded as evidence of a person 'having' the component of ability. This operationalising takes account of the domain of language use relevant to the group being tested.

The operationalised model should then be built on in the drawing up of a blueprint of detailed specifications of 'what the test tests and how it tests it' (Alderson *et al.* (1995: 9). From the point of view of validators, these specifications should explicitly describe what is meant by the ability being tested, how the individual tasks are to elicit evidence of this ability, and how the ability is to be assessed. The methods and procedures that are instrumental in eliciting the evidence of ability should also be laid down in the specifications.

Once the purpose and object of testing have been established, the process of validation can proceed in two ways: by inspection (e.g. seeing how scoring instruments fit a theoretical model of language ability) and by collecting evidence (e.g. raters' scores on sub-tests). These two approaches to validation correspond roughly, but not entirely, to two recognised stages in validation: a priori and a posteriori. A priori validation involves a scrutiny of the test 'as it stands', i.e. before it is put into use, and largely involves inspection. A posteriori validation involves investigating the way the test appears to have worked 'after the event', and largely involves the analysis of scoring data.