Testing the Spoken English of Young Norwegians

A study of test validity and the role of 'smallwords' in contributing to pupils' fluency

Testing the Spoken English of Young Norwegians

A study of test validity and the role of 'smallwords' in contributing to pupils' fluency

Angela Hasselgreen

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge CB2 1RP, UK

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011–4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published 2004

Printed in the United Kingdom at the University Press, Cambridge

Typeface Times 10/12pt. *System* QuarkXPress®

A catalogue record for this book is available from the British Library

ISBN 0521 83613 1 hardback
ISBN 0521 54472 6 paperback

Also in this series:

An investigation into the comparability of two tests of English as a Foreign Language: The Cambridge-TOEFL comparability study
Lyle F. Bachman, F. Davidson, K. Ryan, I.-C. Choi

Test taker characteristics and performance: A structural modeling approach
Antony John Kunnan

Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem
Michael Milanovic, Nick Saville

The development of IELTS: A study of the effect of background knowledge on reading comprehension
Caroline Margaret Clapham

Verbal protocol analysis in language testing research: A handbook
Alison Green

A multilingual glossary of language testing terms
Prepared by ALTE members

Dictionary of language testing
Alan Davies, Annie Brown, Cathie Elder, Kathryn Hill, Tom Lumley, Tim McNamara

Learner strategy use and performance on language tests: A structural equation modelling approach
James Enos Purpura

Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida
Antony John Kunnan

Issues in computer-adaptive testing of reading proficiency
Micheline Chalhoub-Deville

Experimenting with uncertainty: Essays in honour of Alan Davies
A. Brown, C. Elder, N. Iwashita, E. Grove, K. Hill, T. Lumley, K. O'Loughlin, T. McNamara

An empirical investigation of the componentiality of L2 reading in English for academic purposes
Cyril Weir

The equivalence of direct and semi-direct speaking tests
Kieran O'Loughlin

A qualitative approach to the validation of oral language tests
Anne Lazaraton

Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913–2002
Edited by Cyril Weir and Michael Milanovic

European language testing in a global context
Edited by Cyril Weir and Michael Milanovic

Unpublished

A Modular Approach to Testing English Language Skills: The development of the Certificates in English Language Skills (CELS) examinations:
Roger Hawkey

Changing language teaching through language testing: A washback study
Liying Cheng

The Impact of High-Stakes Examinations on Classroom Teaching: A Case Study Using Insights from Testing and Innovation Theory
Dianne Wall

Contents

Acknowledgements

This book – and the study it reports on – could never have proceeded as smoothly as it did without the support of a lot of people. First, I would like to thank my family – close and extended – and my good friends, who always supported me and never complained that I had so little time for them. Two of my sons, Nicholas and John, laboured long on transcribing students' speech, and Nicholas put in sterling work helping assign signals to smallwords. I also owe a great debt of gratitude to Anna-Brita Stenström, first supervisor, then friend, whose astute eyes and ears were available throughout the study, and who gave so much of her own time, so generously. And I must thank Charles Alderson, whose brisk, pertinent e-mail comments opened my eyes to so much that shaped the study. I must also mention Trude Bungum, who sadly died; she made sharing an office a pleasure, and was a true and wise friend. And finally, I would like to thank Sari Luoma for letting me pick her brains during our stay in Lancaster together, in return for the loan of my bike.

Angela Hasselgreen
Bergen, March 2002.

Series Editor's notes

To improve test fairness we need an agenda for reform, which sets out clearly the basic minimum requirements for sound testing practice. Stakeholders in the testing process, in particular students and teachers, need to be able to ask the right questions of any examinations commercial or classroom based. Examination providers should be able and required to provide appropriate evidence in response to these questions.

It is now axiomatic that a test should be constructed on an explicit specification, which addresses both the cognitive and linguistic abilities involved in the language use of interest, as well as the context in which these abilities are to be performed **(theory based validity and content validity)**. A particular administration of a test may fulfil the requirements of both these validities to a greater or lesser extent.

Next in the implementation stage when the test has been administered, we need to look at the data generated and apply statistical analyses to these to tell us the degree to which we can depend on the results **(reliabilities)**.

Finally we can collect data on events after the test has been developed and administered **(concurrent and consequential validities)** to shed further light on the well foundedness of the inferences we are making about underlying abilities on the basis of test results. The focus here is on the value of the test for end users of the information provided and the extent to which such use can be justified. This takes us into the area of **concurrent** validity evidence where a test is measured against other external measures of the construct, and also that of **consequential** validity where the impact of the test on society and individuals is investigated. This consideration implies that validity does not just reside in the test itself or rather in the scores on the test but also in the inferences that are made from them.

In Chapter 2 of this volume Hasselgreen provides a clear exposition of the nature of test validation and offers a comprehensive working framework for the validation of a spoken language test. The reader is also referred to Volume 15 of this series where the operational procedures for test validation adopted by Cambridge ESOL in terms of Validity, Reliability, Impact and Practicality (VRIP) are described. It is interesting to compare the extent to which Hasselgreen's broad conceptualisation of this area matches that of Cambridge ESOL's operationalisation of these VRIP categories. Together they provide a solid grounding for any future work in this area.

In Chapter 3 she examines in detail how communicative language ability (CLA), a central element of a test's theory based validity, might be operationalised in the evaluation of the Norwegian speaking test, for lower secondary school students of English (EVA). As such it represents one of the few reported attempts to operationalise Bachman's seminal cognitive model of language ability.

In Chapters 4 and 5 she takes the broader validation framework developed in Chapter 2 and applies it to the EVA test and so provides test developers with a working example of how validation might be done in practice. She was able to evaluate all aspects of communicative competence in EVA as it had been defined in the literature to date. Published studies of this type are regretfully rare in the testing literature and Hasselgreen's case study illuminates this vital area of our field in an accessible well written account of a validation carried out on this spoken language test in Norway.

Her validation of the existing test system throws up serious problems in the scoring instruments. In particular the band scale relating to fluency does not adequately account for the aspects of CLA measured by the test particularly as regards textual and strategic ability because it lacks explicit reference to the linguistic devices that contribute to fluency. Low inter-rater correlations on *message* and *fluency* discussed in Chapter 5 in the discussion of a posteriori validation based on test scores further points to the problem of vagueness in the existing definitions of these criteria. This provides the link to the second part of the monograph; how to establish 'more specific, unambiguous, data-informed ways of assessing fluency'. As such it addresses the emerging consensus that rating scale development should be data driven.

In Part 2 Hasselgreen accordingly focuses on one aspect of the validation framework that frequently generates much discussion in testing circles, namely how should we develop grounded criteria for assessing fluency in spoken language performance. In Chapter 6 she examines the relationship between small words such as *really*, *I mean* and *oh* and fluency at different levels of ability. According to Hasselgreen such smallwords are present with high frequency in the spoken language and help to keep our speech flowing, although they do not necessarily impact on the content of the message itself. A major contribution of this monograph is the way she locates her argument in relevance theory as the most cohesive way of explaining how smallwords work as a system for effecting fluency by providing prototypical linguistic cues to help in the process of interpreting utterances.

In Chapter 7, based on a large corpus, she reports her research into the extent to which students taking the EVA test used smallwords. She used three groups of students: British native speaker schoolchildren of 14–15 years of age, and a more fluent and a less fluent group of Norwegian schoolchildren of the same age allocated on the basis of global grades in the speaking test. The results support the case that the more 'smallwords' a learner uses, the better

their perceived fluency. Critically she found that the more fluent speakers of English clearly used this body of language more frequently than high and low achieving Norwegian learners, and the range of the words they used was larger especially in turn-internal position to keep going. The more fluent learners used smallwords in a more nativelike way overall and in most turn positions than the less fluent, and also in terms of the variety of forms used and the uses to which they were put. More nativelike quantitities and distribution of smallwords 'appear to go hand in hand with more fluent speech'. The clear implication is that because small words make a significant contribution to fluent speech, such features have an obvious place when developing effective fluency scales. In Chapter 8 she analyses in more detail how students use their smallwords in helping create fluency in communication, what smallwords actually do, providing further corroboration of the findings in Chapter 7.

In Chapter 9 she looks at background variables in relation to small word use such a gender and context, and considers the acquisition of small words. She then looks at the implications of the findings of her research for language education, assessment (task and criteria) and for teaching and learning and in Chapter 10 she summarises her data in relation to the original research questions.

This volume presents the reader with a valuable framework for thinking about test validation and offers a principled methodology for how one might go about developing criteria for assessing spoken language proficiency in a systematic, empirical manner.

Cyril Weir
Michael Milanovic
Cambridge 2004

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.

The place of a priori validation has been recognised increasingly over the past decade. Weir (1988), who regards it as involving 'deliberation on the match between theory and test', reproaches the American literature of the 1980s for underrating the importance of a priori validation in the validation process: 'The concern is much more with the a posteriori relationship between a test and psychological abilities, traits, constructs, it has measured than with what should have been elicited in the first place' (1988: 16). Skehan (1991) calls for a full analytic a priori content validation before testing. Evidence that this is being put into practice is found in research, e.g. that of Shohamy (1994), where both a priori and a posteriori validation are used to examine different facets of a test.

Thus validation can be regarded as involving two distinct stages and approaches. But the question remains of where to look for sources of invalidity which prevent the test from testing what it is supposed to test. Traditionally, guidelines for this investigation have been organised according to types of validation, although, as Messick (1996) puts it, 'validity is now widely viewed as an integral or unified concept' (Messick 1996: 248).

Both for convenience, and to ensure, as far as possible, that no major source of invalidity might slip through the net, the discussion in this section considers validation type by type. Alderson *et al.* (1995) maintain: 'The more different "types" of validity that can be established, the better, and the more evidence that can be gathered for any one "type" of validity, the better' (1995: 171). In the course of the discussion, specific sources of potential invalidity are identified. Construct validity is dealt with last, and provides a systematic way of grouping the potential sources of invalidity so far identified.

The range and number of types of validation identified in the literature varies greatly, with Cummins (1996: 2–3) listing as many as 16, and suggesting even more. However, certain 'core' types of validation, such as *face, content* and *criterion-related*, have been consistently dealt with traditionally, e.g. by Hughes (1989), and, recently, these have normally been accompanied by others, such as *response* and *washback*, e.g. by Alderson *et al.* (1995).

The types discussed in this section are:

- content validation
- face validation
- response validation
- washback validation
- consequential validation
- criterion-related validation
- validation related to reliability
- validation related to test bias
- construct validation.

The main aim of the discussion will be to highlight ways in which validity, in all its aspects, may potentially be violated. In order to keep it relevant to the present research, the discussion will be conducted and exemplified with the direct testing of spoken interaction in mind. By providing an inventory of sources of potential invalidity, a means will be obtained for systematically checking the test, and finding areas where its validity is vulnerable or marred. In order to ensure that this inventory is comprehensive, the chapter concludes with a consideration of validation as a unitary concept, where all the threads unravelled by the discussion are drawn together.

Content validation

Content validation involves checking the test content for what it seems to test, against documentation, such as the specifications, of what it is supposed to test. A major aim in testing students' CLA is typically to make inferences about their ability to cope with the linguistic demands of the wide, non-test domain of 'real life'. For these inferences to be justified, it is necessary that the sample of language collected in the short space of test-time is somehow representative of the language of real-life communication, and relevant to the specified domain. This representativeness is evaluated in the process of content validation, with respect not only to linguistic forms but also to the functions and conditions of speaking.

Content validation is, then, about ensuring that tests get people to do things that are representative of some domain of language use in real life. However, the question remains: 'representative in what way?', and this leads into the issue of authenticity. If the domain is a very restricted one, e.g. in the case of Language for Specific Purposes (LSP) testing, it may be possible to replicate a situation that elicits all the language operations a testee may be expected to perform in real life. In this case, the performance itself is authentic, or representative of real-life communication. This is what Bachman (1990: 301) refers to as 'real-life' (RL) testing.

In most communicative language testing, however, it is not feasible to simulate totally representative performance. Bachman and Palmer (1996) define authenticity as 'the degree of correspondence of the characteristics of a given language test task to the features of a TLU [target language use] task' (1996: 23). The more a test taker is actually doing things that resemble things s/he would normally be expected to do in the target language, then the more readily the scoring of the test performance can be generalised to non-test target language use. S/he will also have a better perception of the test task if it seems relevant to the TLU. However, Bachman and Palmer (1996) also regard 'interactiveness' as an important characteristic of a test task, whereby a test taker's language ability interacts with his/her other characteristics, such as

topical knowledge. While it is impossible to take individual characteristics into account while designing tests, care can and should be taken to avoid favouring groups with certain characteristics. The question of test bias will be taken up on pages 22–23.

Authenticity and interactiveness can be maximised by simulating likely contexts that are representative and relevant to the TLU and by making the need to communicate as genuine as possible, e.g. through using information gaps (see Brown and Yule 1983a). Thus in validating a test's content, representativeness and relevance can be looked for in the kinds of performance to be elicited, and in the kinds of ability that are activated in response to tasks.

Content validation should be carried out by 'experts' (according to Henning 1987, Alderson *et al.* 1995), e.g. on panels. However, Alderson *et al.* warn of the consequences of adverse group dynamics and the two undesirable extremes of total disagreement and agreement by 'cloning', and recommend more systematic, data-driven approaches to content validation. These include collecting ratings on a number of test facets according to given criteria (1995: 174).

Bachman (1990) maintains that content validation must not include simply test items, but also the methods and procedures involved, as these have an effect on the performance elicited, and may prevent the test taker from doing what is specified.

Content validation, as it has been presented here, is clearly closely bound to processes that occur before and during test development. Weir (1988) ascribes a major part of a priori validation to 'matters which relate to content validity' (1988: 17), and much of the a priori validation carried out in the present study will be concerned with issues that threaten content validity. These issues can be summed up as follows:

- faulty or incomplete operationalisation of components of the model of CLA
- poor sampling of the language associated with the underlying theoretical model and domain of CLA, when making tasks
- tasks that do not enable the testees actively to engage their language ability in a reasonably authentic way
- test methods and procedures that may prevent testees from performing in the way intended.

Face validation

According to Hughes (1989: 27), 'a test is said to have face validity if it *looks* as if it measures what it is supposed to measure'. But, as opposed to content validity, which is judged by 'experts', face validity is judged by various groups of non-experts who come into contact with a test.

The status of face validity in language-testing literature is at worst low and at best questionable. Henning (1987) tends to conflate the term with content validity, but differentiates it insofar as 'face validity, unlike content validity, is often determined impressionistically' by, for instance, test takers (1987: 94). Bachman (1990) reinforces the view of face validity as being determined unscientifically by uninitiated people, and poses the questions: 'How do we know who will find what type of test task acceptable?' and 'What do we do if test developers, takers and users disagree?' (1990: 289).

However, it is generally acknowledged that a test that lacks face validity may never be used. And even if it is used, disenchanted test takers will not perform as they should, which will lead to the test's failing on response validity. This will be discussed in the next section.

One criterion of face validity is presented by Alderson *et al*. (1995), who maintain that 'many advocates of CLT [communicative language testing] [...] argue that it is important that a communicative language test should look like something one might do "in the real world" with language.' (1995: 172). Thus the point made in the last section about the importance of authenticity in tasks holds good for face as well as for content validity. Besides authenticity, the degree of familiarity may affect the face validity of a test. Unfamiliar formats and procedures may be rejected out of hand by potential users, or may not be taken seriously by the testees.

Alderson *et al.* (1995) offer suggestions for data-driven research into students' attitudes, tapped after doing or looking at a test. This 'evidence-collecting' process of face validation can thus be either a priori or a posteriori, and should perhaps be as equally concerned with other users, e.g. teachers, as with the test takers' own attitudes. In the case of unfamiliar formats, face validity can be built up, a priori, through a practice testing programme, whereby users are familiarised with samples of material.

To sum up, two factors that may deprive a test of face validity are:

- unfamiliarity of format
- lack of authenticity in the test tasks.

Response validation

In the two previous sections, mention has been made of the response of the test taker to tasks. Content validation places demands of representativeness on the intended response to the test tasks. Face validation plays a part in ensuring that the expected response is forthcoming. The fact is, however, that no guarantee can ever be given that test takers will respond the way they are intended, or expected to. 'The extent to which examinees respond in the manner expected by the test developers' has been termed 'response validity' by Henning (1987: 96), among others.

Henning goes on to suggest that breakdown in response validity is ascribable either to the attitude of the testee in approaching the tasks or to genuine lack of understanding about what to do. A negative attitude may be brought about by tests that lack interest, are demotivating, or are not taken seriously. Lack of understanding can be caused by unfamiliarity with the format or unclear instructions. There is clearly common ground between the factors that cause face and response invalidity.

However, there is more to response validity than a simple acceptance of the test. As far as possible, the person performing the tasks should go through the processes associated with the real-life communication, with a minimum of 'irrelevant' processes. Alderson *et al.* (1995) highlight the current concern for ensuring that testees are actually going through the processes intended while carrying out test tasks:

> an increasingly common aspect of test validation is to gather information on how individuals respond to test items. The processes they go through, the reasoning they engage in when responding, are important indications of what the test is testing, at least for those individuals. (1995: 176)

This information-gathering can only be carried out validly under 'genuine' test conditions, and, as such, belongs to the realm of a posteriori validation. However, as an extension of the a priori process of face validation, opinions can be sought in advance on aspects of the test that contribute to ease of understanding (such as familiarity and clarity of instructions) and motivation (such as degree of appeal and test length).

In sum, response validity may be violated by:

- faulty or incomplete operationalisation of components of the model of CLA
- tasks that do not fully engage the testees in the processes associated with the underlying theoretical model and domain of CLA
- tasks that essentially draw on processes that are irrelevant to the underlying theoretical model of CLA
- tasks that are uninspiring or off-putting and so fail to engage the testee in real communication
- lack of understanding due to unclear instructions or unfamiliarity.

Washback validation

While the beneficial washback of a test might be assessed in terms of its effects on teaching practices and curricula (c.f. Hamp-Lyons 1997: 296), it can only be truly evaluated by what goes on in the learner's mind. Messick (1996) points out that, whatever the effect on teaching practices, positive washback can really only be acknowledged insofar as learning itself is enhanced by the test.

Many factors have been put forward as contributing to the potential washback validity of tests, and these have been summarised by Bailey (1996) as 'the incorporation of:

- language-learning goals
- authenticity
- learner autonomy and self-assessment
- detailed score reporting' (1993: 268).

The first two of these factors have already been touched on in the discussion on content validation, to the extent that it may be assumed that (long-term) language-learning goals and the model of language ability underlying the test are largely compatible (or, at least, should be). The third and fourth factors relate to the test's scoring procedures rather than its tasks. Here the concern is about the way performance is judged and reported, which should carry a clear message about what goes into language ability. And through the involvement of learners themselves in the process, by self-assessment, this message is more likely to penetrate the mind where the learning is taking place, reinforcing the potential washback effect. It should be emphasised, however, that the extent to which these factors are salient depends on the aims and values underlying the testing.

While empirical studies have been carried out on the validation of tests with respect to their washback on educational practices, these tend to be beset with difficulties. A posteriori studies have been known to collapse (e.g. Wall *et al.* 1991) owing to a lack of data on the situation before the testing. And the findings of synchronic studies tend to be muddled by factors such as those relating to individual teaching styles, class sizes or maturity (e.g. Alderson and Hamp-Lyons 1996). The empirical study of the effect of washback on learning is still less feasible. Messick (1996) points out that teacher behaviour being influenced by a test does not imply that learner behaviour is influenced. He maintains, moreover, that any improvement in ability can be ascribed to washback effect only if it can be evidentially linked to the introduction and use of the test, which is extremely difficult to bring about with any confidence.

Messick concludes that we should 'rather than seeking washback as a sign of test validity, seek validity by design as a likely basis for washback' (1996: 252). A valid test, following Messick's argument, will not be subject to 'construct under-representation' or 'construct-irrelevant variance' (1996: 247), both of which would threaten positive washback effect. In other words, washback validity depends, firstly, on all the constructs believed to make up language ability (however defined) being represented in the test, and, secondly, on irrelevant factors such as 'testwiseness' minimally influencing performance and hence scores. The test will therefore carry a comprehensive message of what is to be learnt and will discourage practices that focus on

issues irrelevant to the learning process, such as test technique. Messick's comment: 'for optimal positive washback there should be little if any difference between activities involved in learning the language and activities involved in preparing for the test' (1996: 241–2) might be considered to sum up his message, applying his argument as much to assessment criteria as to test tasks. And as self-assessment is increasingly recognised as an essential learning activity (e.g. in Little and Perclová 2001), an element of self-assessment should ideally be associated with the test (at least in 'practice' versions). It can thus be concluded that washback validity can be threatened specifically by:

- test tasks and methods and assessment criteria that do not fully reflect the model and domain of CLA in an authentic way, or which draw on irrelevant abilities
- scoring procedures that do not fully reflect the model of CLA or which do not encourage the learner to assess his/her own performance.

Consequential validation

'Consequential validity' is used here to denote the extent to which the test results are used in the way intended, and are successful in bringing about the aims of the testing. This validity differs from washback validity in that the latter is largely concerned with the influence of the whole testing process on the teaching and learning situation, whereas the former is concerned with the effects of the end product, i.e. the test result. Bachman and Palmer bring together these influences and effects in what they term the 'impact' of a test 'on society and educational systems and upon those individuals within those systems' (1996: 29). Messick (1995) comments: 'It is ironic that validity theory has paid so little attention over the years to the consequential basis of test validity, because validation practice has long invoked such notions as the functional worth of the test-taking – that is, a concern over how well the test does the job for which it is used' (1995: 744).

Consequential validation is called for by Shohamy (1993) when she states, 'Testers must begin to examine the consequences of the tests they develop. Testers often feel that they have completed their job after obtaining a high reliability and validity and do not find it necessary to observe the actual use of the test' (1993: 37). Messick (1996), in listing 'perennial validity questions', asks: 'Do the scores have utility for the proposed purposes in the applied settings? Are the short- and long-term consequences of score interpretation and use supportive of the general testing aims and are there any adverse side effects?' (1996: 247). Bachman's (1990) warning that, however thoroughly content validation is carried out, the inferences we can draw from the test performance can only be on what a person can do (not what he cannot do)

within the domain of language ability specified (1990: 246), is also salient to consequential validity.

Clearly the most conclusive way of carrying out consequential validation is through a posteriori surveying, e.g. getting test users to respond to salient questions relating to the use and effects of the test scores. Messick (1995) states: 'because performance assessments in education promise potential benefits for teaching and learning, it is important to accrue evidence of positive consequences as well as evidence that negative consequences are minimal' (1995: 746).

However, a priori validation can be carried out by ensuring that the scoring instruments themselves, as well as the procedures for using and interpreting them, potentially 'match' the initial aims of the testing, and by anticipating adverse side effects. In a test such as the EVA speaking test, where the aims are to identify strengths and weaknesses, with a view to adapting future teaching and learning to the specific needs of the individual student, certain particular demands are made of scoring instruments and procedures. Some kind of profiling of the performance should be carried out as part of the procedure, in order to highlight strengths and weaknesses. Band scale descriptors should be explicit in stating what students typically 'can do' at a certain level in order, firstly, to tailor learning to the present state of ability, and, secondly, to give the learner the opportunity to see what s/he needs to be able to do in order to improve. Descriptors should therefore be concrete and positive (see North 1997: 439).

A test of the type being studied here could fall down on consequential validity on account of:

- lack of any analytic feedback on individual strengths and weaknesses
- band-scale descriptors which are vague or negative, so that they do not help learners to realise what they can and need to be able to do
- unclear instructions to users on how (and how not) to interpret test results
- failure to restrict inferences, made from test results, to what the testee can do, in the content domain specified.

Criterion-related validation

So far all the types of validation discussed have involved, to some degree, a priori validation. They have largely concerned what can be done to tests to make sure they will work well. However, in order to find out whether a test actually is working well, test results have to be compared with some external yardstick or criterion. This criterion-related validation can only be carried out a posteriori, using test score data. Two types of criterion-related validation are recognised. The first is 'predictive validation', whereby results of a test whose purpose is to predict future performance are compared with some later

measure of that performance. The second is 'concurrent validation', whereby results of a test of current ability are compared with an external, independent measure of that same ability. Because of the non-predicting nature of the test being validated in the present study, only the concurrent type of criterion-related validation will be considered here.

The first concern in the process of concurrent validation is to find 'a criterion which we believe is also an indicator of the ability being tested' (Bachman 1990: 248). Bachman goes on to cite examples of such criteria: 'the level of ability as defined by group membership, individuals' performance on another test of the ability in question, or their relative success in performing some task that involves this ability' (ibid: 248). Illustrations of these criteria include teachers' estimates, test results on some well-established standardised test and self-assessment.

On a cautionary note, Bachman warns against using measures or estimates of language ability in general as a yardstick for measuring a particular ability. Furthermore, Bachman regrets that criterion-related validation, as a rule, intentionally involves only correlating with measures of the *same* ability, i.e. looking for *convergent evidence.* He believes that it is equally important to look for *discriminant evidence*, correlating test scores with measures of other abilities, with a view to demonstrating, through low correlations, that our test is not significantly measuring these other abilities (1990: 250). Furthermore, he underlines the folly in overreliance on other measures, whose own validity can never be established absolutely. Henning (1987) offers solutions for correcting for unreliability in the other measure, but, as there is no outright 'coefficient of validity' to compute or compare with, we must always regard criterion-related validity as relative to the criterion we choose as our yardstick. In a recent film, where the height of a Welsh mountain was being calculated, based on a survey of surrounding known peaks, the question 'who measured the first mountain?' went unanswered.

Whenever we are measuring something using a yardstick that is not absolute, we need good reason to believe that the yardstick is a sound one. In the case of language ability, one normally dependable source is teachers' judgement, and Alderson *et al.* (1995) present some concrete suggestions for tapping this source. They give the proviso, however, that two teachers should be independently involved, and that either ranking or numerically expressed ratings must be used. Students' self-assessments are also held by many to be a valuable alternative source of assessment (see Oskarrson, 1988). However, Alderson *et al.* (1995: 178) warn that correlations higher than 0.5 to 0.7 are unlikely to be achieved when validating with respect to these non-test external measures.

Shohamy (1994) weakens the case for what we can read into criterion-related validity in her comparison of the 'SOPI' and 'OPI' (respectively Semi-direct and (direct) Oral Proficiency Interview) tests. Here she demonstrates

that two tests of oral language ability, whose results correlate highly, suggesting a mutual external validity, in fact measure abilities that are quite different in some aspects. Shohamy's findings underscore the danger of overreadiness to interpret a positive correlation as 'proof' of criterion-related validity (and the danger of placing too much reliance on a single validation process).

Criterion-related validation is thus a process undertaken as part of a posteriori validation, which gives an indication of how well a test is working. Whatever external measure is used, we must be aware of its limits as a yardstick. The following factors will undermine the value of criterion-related validation:

- using external criteria that measure different abilities from the test in question
- failing to look for discriminant evidence to ensure that the test is not measuring unrelated, irrelevant abilities
- using external criteria whose validity is unknown.

Reliability

Any test score or grade is affected to some extent by factors that have nothing to do with the ability being tested. Some of these factors are totally random, and at its simplest, following *Classical True Score Theory* (Bachman 1990: 184), the reliability of a test score is a way of expressing the 'proportion' of the score that is consistent (over hypothetical retestings), i.e. not based on these random factors. The more recently developed *Item Response Theory* (IRT) (e.g. in Baker 1997), and particularly the use of multi-faceted Rasch analysis, have led to ways of going further than this, identifying and eliminating non-random factors, e.g. those linked to personality or rater bias, that consistently contribute to a score. Ideally, reliability measured this way should tell how much of the score is actually a reflection of the ability being tested.

The relationship between validity and reliability is a finely balanced one. On one hand we have the fact that, as Henning (1987) puts it, 'For most empirical kinds of validity, reliability is a necessary but not sufficient condition for validity to be present' (1987: 89). However, the conflict of interest between validity and reliability is pervasive in the literature. Henning (1987) points out that homogeneity of items, which favours reliability, works against content validity, which depends on diverse and comprehensive items. Skehan (1991) argues that too much significance has been accorded to the 'pursuits of test reliability, item homogeneity and scale unidimensionality', and maintains that 'the pursuit of these goals has been at the expense of the other desirable test qualities. Of these, the most important, by far, is test validity' (1991: 5).

Davies (1990) puts the issue on the line when he says that 'a completely reliable test would measure nothing; and a completely valid test would not measure' (1990: 50). It seems reasonable to conclude that the decision on whether to favour reliability or validity is dependent on the individual test's purpose and its 'stakes'. In a low stakes test of speaking that is to be used in classroom diagnosis, with washback and informative feedback as primary aims, validity should come first. However, reliability clearly cannot be ignored, and causes of unreliability must be identified and eliminated as far as possible.

In a direct test of speaking, where performances are graded according to criteria in band scales, a major source of unreliability lies with the raters' grading. Inter-rater reliability is the extent to which different raters are able to agree on the same performances, while intra-rater reliability is the extent to which the same rater would (hypothetically) be consistent if applying the same criteria to the same performance repeatedly. Investigating rater reliability is an essential part of a posteriori validation. The depth and manner in which this is done will depend largely on the stakes of the test, the availability of raters for experimenting and the degree of responsibility given to individual raters. In a low-stakes test, where classroom teachers will normally be doing the scoring, and where data is taken from non-experimental, 'real-life' scorings, a simple inter-rater correlation at the test piloting stage, using global scores and any sub-scores available, may be the only feasible and worthwhile reliability study to carry out. The information yielded should give some indication of where, in the test procedure, potential sources of rater unreliability lie, and these can then be acted on.

However, a good deal of a priori work can be undertaken to bolster rater reliability. In his discussion on developing more reliable 'band scores', Alderson (1991) mentions several measures which may bring this about. These include the use of profiling with respect to various aspects of performance at different stages in the interaction, the practice of moving from an initial broad band score to a final narrower score, and of course rater training including the use of sample performances. North (1997: 439) emphasises the need for descriptors of performance that are concrete. And the need for clear instructions on the scoring process goes without saying.

Non rating-related threats to reliability in direct testing are less easy to identify through scores, yet can be safeguarded against in advance. These principally concern issues related to test methods and procedures, such as the physical environment of the testing, partner compatibility and the input and instructions given by the tester. Methods and procedures should be clearly defined and consistently adhered to. The possible effect of a weak or dominant partner should be anticipated and counteracted in the test method, and may in fact be investigated, a posteriori, using test score data.

To sum up, the measuring of inter-rater reliability is a necessary part of a posteriori validation, and should give indications of where weaknesses lie.

And measures can be taken a priori to safeguard against many sources of unreliability. Significant sources of unreliability identified in this section as relevant to the current test validation include:

- methods and procedures for testing that are unclear or weakly defined, so that inconsistencies may occur in the way the test is carried out
- the influence of a weak or dominant partner
- band scales or other scoring instruments that are couched in vague terms
- instructions and procedures for scoring that are unclear or weakly defined
- lack of rater training.

Test bias

A test can be said to be biased if a group of testees systematically perform better or worse than the norm, for reasons which cannot plausibly be ascribed to language ability. Bachman (1990) identifies four main categories of sources of test bias:

cultural background

background knowledge

cognitive characteristics

the inclusive category of native language/ethnicity/age/gender.

Cultural background can affect test scores through both test method and test content. Clearly, in a direct test of speaking, regard must be given to the fact that topics that are neutral in certain cultures might be emotive or taboo in others. And the notion of culture itself should be interpreted widely, taking account of the socio-cultural spectrum, the diversity in family set-ups and so on; the test should show sensitivity.

Background knowledge is widely recognised as a potential source of bias. In some test situations, e.g. in LSP testing, this knowledge may be assumed for the group the test is intended for (although the problem can arise as to whether the test reflects degree of language ability or degree of background knowledge). In a test of CLA as such, the range of topics discussable should be specified as part of the domain of CLA. And even having done this, care has to be taken so as not to favour certain groups. Within the topic of 'leisure' for example, teenagers growing up on the coast of Norway should not be assumed to be *au fait* with the niceties of skiing.

Cognitive characteristics are addressed by Bachman (1990), with respect to the effects of field dependence/independence on performance in language tests. Bachman finds that the evidence on which methods are favoured by field dependence or independence is inconclusive, and even contradictory.

Native language, ethnicity, age and gender are all potential sources of test bias which could more plausibly be researched in a study of the present kind. The influence of the L1, for instance, is a widely researched and theorised topic within SLA. Particular native languages have, not unexpectedly, been shown to have a differential influence on foreign language test performance, the 'distance' between the L1 and L2 being particularly telling, as is discussed in Kellerman (1983) and Ringbom (1987).

It is clear from the above discussion that sources of bias lie principally in the format of the actual test – its tasks and procedures followed – although of course it may have an impact on other aspects, such as rating procedures. For this reason test bias will be considered primarily as a threat to content validity.

Where it is feasible, relevant sources of test bias should be investigated as part of a posteriori validation, comparing group scores. However, it is obviously simpler to make comparisons across the few clearly defined groups, e.g. across gender, than with fuzzy or sensitive groupings. What is more, Bachman (1990: 278) makes the cautionary point that group difference in performance may not necessarily be a result of test bias but rather an indication of a difference in the actual language ability of the particular group. In the case of native-language background, this has been shown to be the case (e.g. in Ringbom 1987). And obviously in the case of age, but even in the case of gender where adolescents are concerned, developmental factors cannot be eliminated. Ideally therefore, no conclusions should be drawn on test bias from scores indicating group differences before checking with other across-group measures of the same ability.

With this in mind, it seems that while individual a posteriori studies may be carried out on each of these potential sources of bias, a priori work done to prevent bias may be the most satisfactory and comprehensive way of validating tests in this respect. This can be done by developing and discussing the test with all the relevant potential causes of bias in mind. Major causes of bias can be identified as:

- cultural background
- background knowledge
- native language
- ethnicity
- age
- gender.

Construct validation

'The concept of construct validity has been widely agreed upon as *the* single fundamental principle that subsumes various other aspects of validation' (Cummins 1996: 5). Because of this, all the aspects of validation discussed so

far can be brought together under the umbrella of construct validation, and so it has a natural place at the end of this discussion on validation. However, it would have been equally logical to place construct validation at the beginning of the discussion, seeing that it concerns issues that are so fundamental that, unless they are valid, all other types of validation are meaningless. These two perspectives on construct validity exist because it has two interpretations – one wide and the other narrow.

Construct validity, as its name suggests, has to do with 'constructs', which, according to Bachman (1990), 'can be viewed as definitions of abilities that permit us to state specific hypotheses about how these abilities are or are not related to other abilities, and about the relationship between these abilities and observed behaviour' (1990: 255). We are drawing on hypotheses of this nature when we issue a test result that describes ability, based on evidence in performance, which is interpreted in the light of a theoretical model of ability. In construct validation, we put these hypotheses to the test.

Clearly, anything that interferes with the chain of processes that takes us from an underlying model of ability to a test result will weaken these hypotheses, and thus, in the broader view, construct validation must include all types of validation. We will return to this broad view later in this section. Meanwhile, it is necessary to take a narrower look at construct validation, in order to test the hypotheses that are, to some extent, taken for granted when carrying out other forms of validation.

The narrower view of construct validity

In the discussion on content validation, pages 12 – 13, it was claimed that we must ensure that the sample of language elicited in a test is representative of the language behaviour that we have defined as 'evidence' of CLA. The assumption is implied that the way components of CLA have been operationalised, prior to test development, is dependable (otherwise we would not be able to recognise evidence of ability correctly).

Moreover, certain assumptions are made about the scoring instruments and the way they work. Firstly, we assume that testees are likely to score differently on the various component measures (otherwise these measures would have no point). Secondly, we trust that the way constructs are clustered within the individual sub-scales reflects actual affinities in abilities (otherwise they would be unusable). Moreover, we assume that the descriptors of abilities contained in the different levels of band scale present a more or less true picture of students' ability at these levels (otherwise they would give misleading information).

Three central hypotheses can thus be identified as being made, concerning the type of testing being studied here:

- components of CLA are operationalised in a dependable way
- the way constructs of ability are clustered and differentiated in scoring reflects actual affinities and divisions of ability within CLA
- the descriptions of performances at different levels in band scales present a picture of progression through levels more or less as it actually occurs.

These hypotheses should be tested, as far as is possible, in construct validation, through both 'logical analysis and empirical investigation' (Bachman 1990: 256). The first hypothesis can best be tested a priori, by examining the way components of CLA have been operationalised in the recent literature, bearing in mind the situation (i.e. domain of language use) of the particular group being tested.

The second hypothesis, concerning the clustering of constructs of ability, can largely be tested a priori, by reference to literature and testing convention, since this issue is not entirely unique to a particular testing context. It should, preferably, also be tested a posteriori by the use of empirical analysis. Alderson *et al.* (1995) suggest various means of carrying out this empirical analysis, including internal correlations between scores on sub-skills, comparisons with biodata (age, years of study, etc.) and factor analysis of sub-skill scores, all of which methods combine quantitative data analysis with qualitative theory-based interpretation.

The third hypothesis concerns the description of ability or performance at different levels, which is unique to the particular testing context. This should therefore be tested a posteriori, e.g. by using raters' scores on separate aspects of performance of testees at different (global) levels, or by examining transcripts of actual performance by testees at different levels. Pilot investigations can be carried out a priori to assist in the initial development of band-scale descriptors, which should preferably draw on SLA research and theory when possible.

To sum up, construct validity in its narrow sense can be undermined by:

- faulty or incomplete operationalisation of components of the model of CLA
- clustering and division of constructs in the scoring system that are not supported by primary or secondary empirical evidence
- the creation of band-scale descriptors of performance at different levels of ability, which are not supported by some primary empirical evidence.

The broader, unifying, view of construct validity

In its broader sense, construct validity can be regarded as constituting the fundamental, unifying, issue in test validation. Messick (1995) states:

> Indeed, validity is broadly defined as nothing less than an evaluative summary of both the evidence for and the actual – as well as the potential

> – consequences of score interpretation and use (i.e. construct validity conceived comprehensively). This comprehensive view of validity integrates considerations of content, criteria and consequences into a comprehensive framework for empirically testing rational hypotheses about score meaning and utility. (1995: 742)

Messick goes on to reiterate that two major threats to construct validity are construct under-representation and construct irrelevancies. Either we may not be fully assessing the constructs that we claim to be assessing (throughout the process, from elicitation of evidence to interpreting and acting on scores), or we may be allowing other, irrelevant, constructs or abilities to affect scores and hence their consequences.

In order to safeguard against these threats, many questions need to be addressed: we need to know that the test tasks primarily engage the construct or ability we are assessing, in the way it is generally engaged in a target language use (TLU) situation; we need to know that the way information is presented in test results reflects the way the construct is believed to be made up; we need to know that we can depend on scores not fluctuating owing to irrelevant factors and that these scores are generalisable to non-test performance; we need confirmation that the test scores seem to apply to the ability intended, through comparison with other measures of the same or different abilities; and, finally, we need to know that whatever scores are saying about the construct is being interpreted and acted on in the way intended.

Thus, while Messick (1996) upholds the view of validity as a unified concept, he maintains:

> this does not imply answering only one overarching question or even several questions separately or one at a time. Rather it implies an integration of multiple complementary forms of convergent and discriminant evidence to answer an interdependent set of questions [...]
>
> In particular, six distinguishable aspects of construct validity are highlighted as a means of addressing central issues implicit in the notion of validity as a unified concept. These are content, substantive, structural, generalizability, external and consequential aspects of construct validity. In effect, these six aspects function as general validity criteria or standards for all educational or psychological measurement.
>
> (1996: 248)

Messick's view of validity as a unified concept provides a framework for systematically examining the overall validity of a test, by taking each of his aspects in turn, and considering what may threaten that particular aspect. This framework is outlined in 'A unified framework for validation' pages 28 – 31. However, before moving on to this, it is worth summarising all the threats to validity that have been uncovered so far.

Threats to validity summarised

The threats to validity associated with the generally discussed 'types' of validation can be summarised as follows:

CONTENT VALIDATION

- faulty or incomplete operationalisation of components of the model of CLA
- poor sampling of the model and domain of CLA when making tasks
- tasks that do not enable the testees to engage their language ability actively in a reasonably authentic way
- test methods and procedures that may prevent testees from performing in the way intended.

FACE VALIDATION

- unfamiliarity of format
- lack of authenticity in the test tasks.

RESPONSE VALIDATION

- faulty or incomplete operationalisation of components of the model of CLA
- tasks that do not fully engage the testees in the processes associated with the underlying theoretical model and domain of CLA
- tasks that essentially draw on processes that are irrelevant to the underlying theoretical model of CLA
- tasks that are uninspiring or off-putting and so fail to engage the testee in real communication
- lack of understanding due to unclear instructions or unfamiliarity
- failure to restrict inferences, made from test results, to what the testee can do, in the content domain specified.

WASHBACK VALIDATION

- test tasks and methods and assessment criteria that do not fully reflect the model and domain of CLA in an authentic way, or which draw on irrelevant abilities
- scoring procedures that do not fully reflect the model of CLA or which do not encourage the learner to assess his/her own performance.

CONSEQUENTIAL VALIDATION

- lack of feedback on individual strengths and weaknesses
- band-scale descriptors that are vague or negative, so that they do not help learners to realise what they can and need to be able to do
- unclear instructions to users on how (and how not) to interpret test results.
- failure to restrict inferences, made from test results, to what the testee can do, in the content domain specified.

CRITERION-RELATED VALIDATION

- using external criteria that measure different abilities from the test in question
- failing to look for discriminant evidence to ensure that the test is not measuring unrelated, irrelevant abilities
- using external criteria whose validity is unknown.

RELIABILITY

- methods and procedures for testing that are unclear or weakly defined, so that inconsistencies may occur in the way the test is carried out
- the influence of a weak or dominant partner
- band scales or other scoring instruments that are couched in vague terms
- instructions and procedures for scoring that are unclear or weakly defined
- lack of rater training.

TEST BIAS with respect to:

- cultural background
- background knowledge
- native language
- ethnicity
- age
- gender.

CONSTRUCT VALIDATION

- faulty or incomplete operationalisation of components of the model of CLA
- clustering and division of constructs in the scoring system that are not supported by primary or secondary empirical evidence
- the creation of band-scale descriptors of performance at different levels of ability, which are not supported by some primary empirical evidence.

A unified framework for validation

The six aspects of construct validity cited by Messick (1995, 1996) are used here as the basis of a framework for validation as a unified concept, which takes into account all the potential threats to validity exposed in this section so far. The explanations of the six aspects, given below, are intended as interpretations rather than as paraphrases. However, in personal correspondence (1998), Messick has commented that this formulation is clearly in the spirit of what he intended. What is most important, in the context of this book, is that the framework provides a way of housing all the sources of invalidity identified here, and that some structure is obtained which

provides a theory-based rationale for the way the process of validation can be carried out, as well as a systematic way of drawing conclusions on test validity. By theory-based, it is meant that the validation process is founded on a single underlying theory, as embodied in Messick's framework, of what validity consists of. This can be contrasted with validation processes that look individually at conventionally accepted types of validity, such as face validity and content validity, without linking them to a unified abstraction of validity.

As Messick (1995) puts it: 'The six aspects of construct validity afford a means of checking that the theoretical rationale or persuasive argument, linking the evidence to the inferences drawn, touches the important bases' (1995: 747).

Six central aspects of validity

Messick's (1995, 1996) six central aspects of validity are interpreted broadly as follows:

The CONTENT aspect of validity concerns the extent to which the test's design and technical quality enable its tasks to elicit language products (i.e. things people say) that contain representative and relevant evidence of ability (with respect to 'normal' TLU (target language use)) .

The SUBSTANTIVE aspect of validity concerns the extent to which testees actually go through processes (i.e. things people do) associated with the underlying theoretical construct of language ability, taking into account the domain of TLU. Furthermore, it concerns the degree to which the variance in test scores is accountable by the way these processes are carried out.

The STRUCTURAL aspect of validity concerns the extent to which the structure of the scoring procedure, e.g. in band scales, actually reflects the structure of language ability as it is theoretically portrayed or empirically shown to be composed.

The GENERALISABILITY aspect of validity concerns the extent to which the test score/assessment of a person's ability matches the way that person's ability would generally be assessed (on the same criteria), whether in another taking of the test or in a non-test TLU. Thus generalisability involves internal consistency of test scores (reliability), as well as the extent that they can be taken generally to apply to performance in TLU (which concerns the CONTENT, SUBSTANTIVE, STRUCTURAL and EXTERNAL aspects of validity).

The EXTERNAL aspect of validity concerns the extent to which the empirical relationship with other assessments of the same ability (or lack of relationship, in the case of other abilities) supports the way the score is meant to be interpreted.

The CONSEQUENTIAL aspect of validity concerns the way scores are interpreted and acted upon, and the after-effects of the testing, e.g. on teaching and learning. This includes 'washback', i.e. the test's impact on activities carried out in preparation for the test (which concerns the CONTENT, SUBSTANTIVE and STRUCTURAL aspects of validity).

A validation framework

The framework outlined here provides an overview of all the sources of potential invalidity identified in this chapter, and at the same time gives an indication of how these combine to threaten each of the six central aspect of validity. It is not intended as a rigidly partitioned taxonomy, which would run counter to the notion of validation as a unified concept. There is, as Messick (1996) suggests, an interdependence in the factors that contribute to validity, so that what poses a threat to any one aspect may well threaten another. Because there is some explicit overlap in what has been found to threaten certain aspects of validity, it has been found convenient to subsume some aspects under others, as will be indicated. Moreover, the factor listed first* under the content aspect is so fundamental that it can be regarded as a threat to any aspect of validity.

The CONTENT aspect of validity may be threatened by:

- faulty or incomplete operationalisation of components of the model of CLA*
- poor sampling of the language associated with the underlying theoretical model and domain of CLA, when making tasks
- unclear instructions or unfamiliarity of format (which prevent tasks from being done as intended).
- test methods and procedures that may prevent testees from performing in the way intended.
- test bias associated with cultural background, background knowledge, native language, ethnicity, age, or gender.

The SUBSTANTIVE aspect of validity may be threatened by:

- tasks that do not fully engage the testees in the processes associated with the underlying theoretical model and domain of CLA
- tasks that essentially draw on processes that are irrelevant to the underlying theoretical model of CLA
- tasks that do not enable the testees to actively engage their language ability in a reasonably authentic way
- tasks that are uninspiring or off-putting and so fail to engage the testee in real communication.

The STRUCTURAL aspect of validity may be threatened by:

- scoring procedures that do not fully reflect the specified model and domain of CLA
- clustering and division of constructs in the scoring system that is not supported by primary or secondary empirical evidence
- the creation of band-scale descriptors of performance at different levels of ability which are not supported by some primary empirical evidence.

The GENERALISABILITY aspect of validity may be threatened by:

- methods and procedures for testing that are unclear or weakly defined, so that inconsistencies may occur in the way the test is carried out
- scales or other scoring instruments that are couched in vague terms (and so give rise to different rater interpretations)
- instructions and procedures for scoring that are unclear or weakly defined
- lack of rater training (so that scoring is potentially inconsistent)
- the influence of a weak or dominant partner

as well as lack of CONTENT, SUBSTANTIVE, STRUCTURAL and EXTERNAL validity.

The EXTERNAL aspect of validity may be threatened by:

- using external criteria that measure different abilities from the test in question
- failing to look for discriminant evidence to ensure that the test is not measuring unrelated abilities
- using external criteria whose validity is unknown.

The CONSEQUENTIAL aspect of validity may be threatened by:

- test tasks and methods that draw on irrelevant abilities
- scoring procedures that do not encourage the learner to assess his/her own performance
- lack of any analytic feedback on individual strengths and weaknesses
- band-scale descriptors that are vague or negative, so that they do not help learners to realise what they can and need to be able to do
- unclear instructions to users on how (and how not) to interpret test results
- failure to restrict inferences, made from test results, to what the testee can do, in the content domain specified

as well as lack of CONTENT, SUBSTANTIVE and STRUCTURAL validity.

Towards the validation process

The above framework forms the structure for the validation process of the EVA test of speaking, which is described in the continuation of Part One, first a priori then a posteriori. In the a priori validation, each of the six aspects of validity is examined in turn, taking into account all potential sources of invalidity in the test as it stands. Conclusions are then drawn on what appear to be – prior to testing – the relative strengths of the different aspects of validity, and on which aspects seem most in need of further investigation. In the a posteriori validation, test scores are analysed to see the extent to which some of these conclusions are borne out in the actual testing. This, in turn, highlights certain aspects as being in need of deeper investigation, described in Part Two.

As a preliminary stage in the a priori validation process, it is necessary to address the fundamental questions of what constitutes the model and domain of CLA underlying the EVA speaking test, and how the components in this model are operationalised so that they can be recognised in students' language. As has been emphasised, this process of defining and operationalising is fundamental in affecting every aspect of validity, and is thus made the subject of the next chapter.

3 Communicative language ability

When we test, we test something. If 'does the test test what it is supposed to test?' holds the title of the most important question in all language testing, as Alderson *et al.* (1995: 170) claim, the question 'what are we supposed to be testing?' must be a close runner-up. The 'thing' that is supposed to be tested in the EVA speaking test is 14–15-year-old Norwegian students' communicative language ability (CLA) in English spoken interaction. Prior to developing a test of spoken CLA, one of the first tasks is to decide what CLA, in the case of the testees, consists of and how it can be recognised in their speaking. This chapter recounts the logical process of this decision-making, as it was carried out, prior to developing the EVA speaking test. This process begins with theoretical considerations of the constructs making up communicative language competence and ends with a concrete description – or operationalisation – of what students should be able to do with their language in spoken communication.

The notion of 'communicative competence' was introduced by Hymes (1972), who rejected the (then) Chomskyan, strongly grammatical, characterisation of language competence (see Chomsky 1965), pointing out that it failed to accommodate socio-cultural knowledge, essential to appropriate language use in actual communication. Furthermore, Hymes (1972: 282) proposed that the notion of ability *to use* language be incorporated alongside knowledge as part of communicative competence. The complexity of the relationship between *competence*, underlying *knowledge* of language and the *ability* to put this knowledge into use in communication will not be fully entered into here, but it is necessary to clarify how these terms will be used and related in this chapter.

According to the Norwegian school curriculum, M87 (valid at the time of test development), students are primarily expected to show that they have the ability to use their English to communicate, rather than to exhibit knowledge of the metalanguage or of prescribed topics. A test of speaking in this context should therefore be primarily concerned with eliciting and assessing what students are able to do with their language, i.e. their CLA. In order to test this ability it must be operationalised, i.e. described in terms of concrete behaviour, so that tasks can be designed to elicit it and so that it may be recognised in the test performance.

However, the ability to use language is dependent on having knowledge about the language. And, to use Hymesian terminology, the combination of underlying knowledge about a language and the ability to use it make up communicative competence. Before attempting to describe CLA in concrete terms, therefore, consideration needs to be given to what has been theoretically ascribed to communicative competence, which involves both language knowledge and the ability to put it into practice.

The operationalisation of CLA thus involves several stages. First, a model of communicative competence, in terms of its different components, should be defined theoretically. Next, the domain of language use must be identified, defined by the situations or conditions under which students are likely to be communicating through the medium of speaking. Thirdly, it must be decided how each component of communicative competence is actually likely to be manifested in this domain of use, i.e. what students might be expected to be able to do with their language knowledge in real communication.

The components of CLA operationalised in this process provide a model, or blueprint, for making test specifications. It must be emphasised from the outset that no claim is made that components of ability are independent or that they can be assessed individually. These components are in fact generally acknowledged to be interdependent (e.g. by Savignon 1983: 46), and the subject of whether any aspect of performance can be separately assessed will be returned to in Chapter 4 pages 58 – 95. However, it is important that the language elicited in testing provides evidence of all sides of language ability, and that the statements made as a result of the test are wide-ranging. A comprehensive model of CLA safeguards against the omission of essential parts of this ability in the design of either test tasks or rating procedures.

Towards a model of communicative competence

Hymes (1972) presents his concept of communicative competence in a seminal paper relating primarily to the language development of disadvantaged children, in which he reacts to the inadequacy of the hitherto grammatical, idealised, perspective on language competence. Hymes calls for a system of describing this competence that takes account of, firstly, knowing not only what it is possible to say in a language but also what is feasible, appropriate and probable in the speech community, and secondly, being able to use this knowledge in communicating in a variety of specific, real situations. Only by reference to such a system does he feel that a fair judgement of the competence of language users can be reached.

In the ensuing decades, many attempts have been made to propose and further develop such a system, or model, of the communicative competence of second- (used here to include ‘foreign-’) language learners, with milestones

provided by, for example, Canale and Swain (1980), Canale (1983) and Bachman (1990). However, the immediate concerns in the works produced on the subject vary in their focus. Canale and Swain (1980) set out to produce a set of 'guidelines in terms of which communicative approaches to second language teaching methodologies and assessment instruments may be organised and developed' (1980:1). Tarone and Yule (1989) analyse communicative competence in the interest of establishing 'what learners should know' (1989: 67). Bachman, on the other hand, is acting primarily in the interest of language testers, in his attempt to identify, with empirical support, measurable constructs of CLA that are independent, as far as this is possible. The DBP (Development of Bilingual Proficiency) Project cited by Schachter (1990) has as its primary goal finding empirical support for a hypothesised three-trait model of language proficiency (1990: 46).

Thus, it seems that a model of language competence (or ability), like a language test, should be designed with a purpose in mind. A model of communicative competence used as a basis for the current testing must be suited to describing what makes up communicative competence comprehensively enough to cover the essentials of this competence, yet compactly enough to be fully reflected in test tasks and scoring instruments. With this in mind, the quest for a suitable model of CLA begins with a review of some of the major models of communicative competence.

Models of communicative competence reviewed

Canale and Swain (1980) build largely on Hymes (1972) to produce a tangible, three-component model of communicative competence. They accept, in principle, Hymes' notion that the characteristics of 'speech events', such as participants and setting, determine our choice of language form. They interpret Hymes as proposing a four-component view of communicative competence: the interaction of *grammatical, psycholinguistic, sociolinguistic* and *probabilistic* systems of competence (c.f. Canale and Swain, 1980: 16). They believe that grammatical and sociolinguistic competence are equally essential components of competence, and, while not adopting in their proposed model Hymes' psycholinguistic component, they maintain that his fourth, probabilistic, component of competence, whereby the learner has a feeling for the likelihood of a form being used, is present throughout their proposed model as a sub-component. Perhaps the most radical contribution they make, however, is the inclusion of *strategic* competence, which, they maintain, 'speakers employ to handle breakdowns in communication', and 'to cope in an authentic communicative situation and [...] to keep the communicative channel open' (1980:25).

Canale and Swain's 1980 model for CLA can be condensed into three major components:

- *grammatical competence* (knowing and using lexis, morphology, syntax and phonology);
- *sociolinguistic competence* (knowing and using rules for appropriate language use, and rules of discourse);
- *strategic competence* (being able to compensate and cope in the face of shortcomings and potential breakdowns).

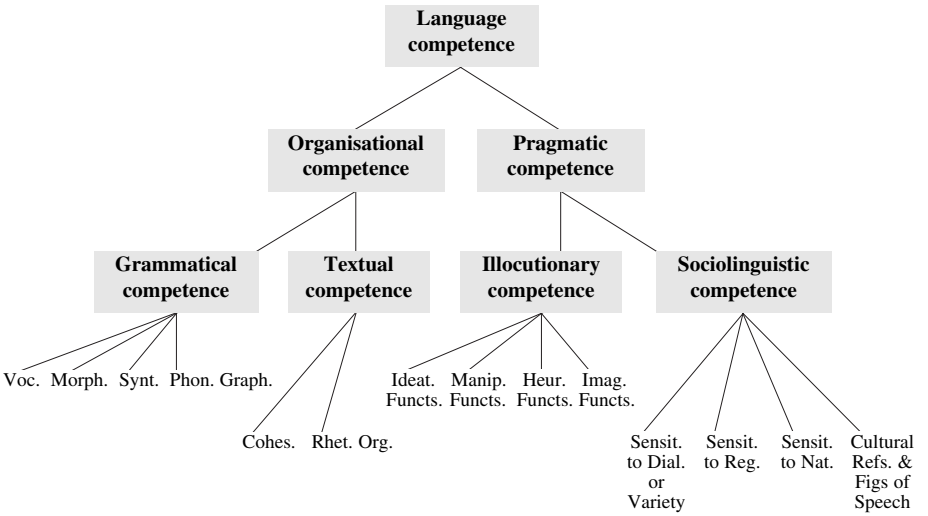
In 1983, Canale modifies this 1980 model of communicative competence, principally by the extraction of the knowledge of discourse rules from within sociolinguistic competence, to comprise an independent component (1983: 9):

- *discourse competence* (knowing how to combine forms and meanings into unified spoken and written texts in different genres).

Savignon (1983) presents a four-part model of communicative competence with the same components as in Canale's (1983) model, while elaborating on certain components. In her explanation of sociolinguistic competence, she defines the social context as including the roles of the participants and the information they share, as well as the function of the interaction, maintaining that 'Only in a full context of this kind can judgements be made on the appropriateness of a particular utterance in the terms elaborated by Hymes'(1983: 37). And strategic competence, Savignon maintains, helps us to cope when faced with questions of the type 'What do you do when you cannot think of a word? What are the ways of keeping a channel open while you pause to collect your thoughts?....' (1983: 40). As almost any sustained piece of communication will testify, whether between native or non-native speakers of a language, the ability to cope with such 'potential breakdown' situations is essential, and, as Savignon puts it, 'distinguishes highly competent communicators from those who are less so' (1983: 40).

Bachman (1990) presents a model of communicative language ability (CLA) in which strategic competence is kept distinct from language competence. This model has four components: *grammatical, textual, illocutionary* and *sociolinguistic*, as shown in Figure 3.1. These components are grouped under two superordinate components: *organisational* and *pragmatic*, which, Bachman maintains, are shown by factor analysis to be independent insofar as individual learners are likely to exhibit unrelated levels in these two competences. However, he emphasises that they are not to be regarded as isolated and independent, maintaining that they react with each other and with other features of the context in language use (1990: 86).

Figure 3.1: Components of language competence (from Bachman 1990: 87).



Bachman's description of the four competences builds on what has already been described in the earlier models discussed here. Grammatical competence includes 'a number of relatively independent competences such as the knowledge of vocabulary, morphology, syntax and phonology/graphology' (1990: 87). Textual competence, corresponding broadly to what Savignon termed 'discourse competence', is defined as including 'the knowledge of the conventions for joining utterances together to form a text' (1990: 88). Bachman goes further than his predecessors, in extending textual competence to cover the ability to 'organise and perform turns in conversational discourse' (1990: 88).

Under the umbrella of 'pragmatic competence', Bachman refers, firstly, to knowing how to perform and interpret the illocutionary force of a speech act (c.f. Searle 1969) according to the conventions of the speech community, terming this 'illocutionary competence'. The functions a language user should be able to perform, through the illocutionary forces of speech acts, are subdivided into four categories, taken from Halliday and Hasan (1976): ideational, manipulative, heuristic and imaginary.

Secondly, pragmatic competence requires that the language user be able to choose an appropriate, or acceptable, language form, in line with the social or discoursal situation that pertains. Bachman terms this ability 'sociolinguistic competence', subdivided into sensitivity to dialect or variety, register and naturalness, and competence in using cultural references and figures of speech.

In a more recent treatment of language ability, Bachman and Palmer (1996) preserve these four components, terming them *grammatical* and *textual* (organisational), *functional* and *sociolinguistic* (pragmatic) *knowledge*. Together with *strategic competence*, these are regarded as components of language ability, which interact with personal characteristics and topical knowledge to make up the 'characteristics that individual language users' bring to the particular situation (1996: 62). The implication of the fact that the ability to put language into practice is partially dependent on personal characteristics and topical knowledge is that, in testing, care must be taken to ensure that the effect of these latter attributes should be either minimised as far as possible in order that a true measure of language ability is achieved, or used positively to facilitate this. This underscores the importance of checking for possible test bias, as discussed on pages 22 – 23.

Bachman and Palmer expand on the separation of strategic competence, which they conceive of as 'a set of metacognitive components or strategies, which can be thought of as higher order executive processes ...'(1996: 70). This view of strategic competence as processes (specifically goal setting, assessment and planning) common to all language use is quite different from the more product-oriented view of Savignon (1983), who perceives strategic competence as realisable by specific types of language use.

While the impression may have been given in the discussion so far that models of communicative competence are ever expanding in the number of their components, or that the existence of independently testable components is established, these impressions are not universally supported, as is illustrated by Schachter (1990) in her critique of the validation carried out by the Development of Bilingual Proficiency (DBP) Project during the 1980s. The project used a three-trait model of communicative competence based largely on Canale and Swain's (1980) and Canale's (1983) models, and, in its validation study, attempted, in vain, to corroborate the independent existence of these traits by factor analysis.

The traits constituting the model were *grammatical, discourse* and *sociolinguistic competence.* Schachter's criticism of this model is largely based on the lack of clear distinction made between the latter two components, both on the part of those involved in the DBP Project and in the way the components have been characterised in the then recent literature. Drawing on several current definitions of 'discourse competence', Schachter maintains:

> If 'discourse competence' is to be viewed as knowledge of the structure of text (in the larger sense of the meaning of text, subsuming both written and oral text), then it would be more appropriate to view it as part of sociolinguistic competence. (1990: 42)

Schachter believes that a 'knowledge of text' cannot ignore such elements as the role of participants or the purpose of the interaction, which are normally

looked on as belonging to the domain of sociolinguistics. Moreover, sociolinguistic competence involves the ability to make appropriate choices of language according to the narrower context of the discourse itself as well as to the culture in which the discourse occurs. In other words, this competence encroaches onto the territory of discourse competence. In short, Schachter sees little justification in maintaining that these two competences exist distinctly. She proposes that communicative competence may be composed of 'two kinds of competence – grammatical and pragmatic – and that sociological phenomena interact with these two components at all levels' (1990: 44).

Schachter concedes that the DBP research has practical educational value through what it contributes to the descriptive knowledge of what communicative competence is. However, her conclusions must be construed as a warning to those attempting to make and validate tests that claim to measure separate traits, based on uncorroborated theoretical models of communicative competence.

To sum up, the development of models for communicative competence over the past two decades or so is a mixed story of stability and flux. The supplementing by Hymes (1972) of what might be termed grammatical competence with something akin to sociolinguistic competence has been preserved in subsequent models. The third and fourth competences – discourse/textual and strategic – were established in the 1980s by Canale and Swain and by Savignon. In the 1990s, Bachman has chosen to place strategic competence outside language competence, regarding it as a processing competence, interacting with all components of language competence or knowledge to bring about ability to use language. Moreover he has extracted from sociolinguistic competence a fifth competence – illocutionary competence – which is concerned with the expression and interpretation of speech acts. Schachter (1990), on the other hand, has collapsed the model somewhat, questioning whether there is any justification in separating discourse and sociolinguistic competence.

And while some rather tentative claims are cited as having been made about the existence of separately measurable components, these tend to involve no more than two or three components, and findings are inconclusive. There does, however, seem to be a general consensus that components are not isolated and that they interact with each other and with other features in the context of language use.

A suitable model of CLA

Having considered the components assigned to communicative competence in the models of some leading researchers, and taking into account the purposes their models were designed for, the next step involves deciding which

components to include in a suitable model of communicative language ability (CLA) for the current, EVA, test situation. This model is described as one of *ability*, rather than *competence,* since it has primarily to serve as a basis for describing the actual language behaviour that indicates what students are able to do with their language. However, no attempt is made to measure any component of ability in isolation. This means that, while the model cannot omit any essential part of ability, it does not need to be discretely partitioned; i.e. 'overlap' is acceptable. These factors having been taken into account, the following four-part model of CLA is proposed:

- MICROLINGUISTIC (to replace the term 'grammatical') ABILITY
- TEXTUAL ABILITY
- PRAGMATIC ABILITY
- STRATEGIC ABILITY.

MICROLINGUISTIC ABILITY

Microlinguistic ability (using the name adapted from Weir (1993: 31)) is included in the model because it is the most stable of all components typically found in models of communicative competence. There seems to be widespread agreement that there is a body of knowledge, essential for any communication, which is generally rule-governed and (relatively) independent of the environment of communication. The term 'microlinguistic' is associated with such strictly linguistic areas as phonology, morphology and syntax (Crystal 1991: 219) and 'units of language at the level of the sentence and below (phrase, word morpheme)' (Wales 1989: 229). It is therefore preferred to 'grammatical', being less exclusive, and covering more precisely the areas assigned to it here (vocabulary, morphology, syntax and phonology). Microlinguistic ability can be briefly defined as being able to access a knowledge of the essential systems of language at the level of the sentence or utterance (defined as 'the physical realisation of a sentence' in Wales 1989) and below.

TEXTUAL ABILITY

Textual ability is included as a component in its own right, following Bachman (1990, 1996). It can be regarded as a descendant of the component referred to as 'discourse competence' by Canale (1983) and by Savignon (1983), who maintains that it involves 'the connection of a series of utterances to form a meaningful whole' (1983: 38). Schachter (c.f. 'Models of communicative competence reviewed' pages 35 – 39) prefers not to separate discourse from sociolinguistic competence, regarding the difference as being largely a matter of how widely the notion of 'text' is interpreted. However, her objection is primarily to regarding these as independently measurable traits, rather than as two recognisable elements of competence.

Support for maintaining a distinction between these two types of ability – *textual* and (here) *pragmatic* – is drawn from Halliday and Hasan (1976), who suggest that two categories of 'things' are involved in making a text coherent. A text, they state, 'has texture, and this is what distinguishes it from something that is not a text. It derives this texture from the fact that it functions as a unity with respect to its environment' (1976: 2). Their definition of a text reflects what they regard as external and internal aspects of texture: 'A text is a passage of discourse which is coherent in these two regards: it is coherent with respect to the context of situation, and therefore consistent in register; and it is coherent with respect to itself, and therefore cohesive' (1976: 23). The authors' comment that 'Neither one of these two conditions is sufficient without the other, nor does the one by necessity entail the other' (1976: 23) clearly suggests that learners' being able to produce or interpret discourse coherently is contingent on their being able to satisfy *both* of these conditions of texture. Thus there appear to be grounds for recognising two components of ability which together enable learners to give 'texture', in its widest sense, to their discourse.

The second of these components, concerning the ability to make a text coherent with respect to itself, is what is referred to here as textual ability. It involves being able to create cohesion in a text, not only in Halliday and Hasan's (1976) explicit sense of cohesion as the expression of semantic relations, but also in Bachman's (1990) sense, with respect to the ability to use the markers and routines that build structure into conversation, as well as the organisational ability to structure information.

PRAGMATIC ABILITY

Pragmatic ability, to use Bachman's (1990) terminology, can be regarded as the ability to make discourse coherent 'with respect to the context of situation' (Halliday and Hasan 1976: 23). This ability is taken here to approximate to what has been called 'sociolinguistic competence' in all the models so far discussed, while taking into account what Bachman termed 'illocutionary competence'. Without pragmatic ability we can probably get across the literal message of what we want to communicate. However, our interaction will not be 'typical' (Mey 1993: 49).

The term 'pragmatic' is chosen to reflect the wide scope of this component of ability as it is understood here. Mey (1993) offers as a possible definition of pragmatics: 'the study of the conditions of human language uses as these are determined by the context of society' (1993: 42). He goes on to distinguish between two elements in this context: the *societal* – referring to what is determined by 'society' – and the *social* – determined by the circumstances surrounding the interaction itself. The term pragmatic ability, as it is used here, reflects these two elements and concerns the ability to use and interpret language in the way that it is typically used and interpreted by the society and in the particular situation in which the communication is taking place.

STRATEGIC ABILITY

Strategic ability is included as a free-standing component in this model of CLA, principally because the model is to be deployed in a test of speaking. Bachman and Palmer (1996: 71) identify three sets of metacognitive strategies as making up strategic competence: *goal setting* (deciding what one is going to do), *assessment* (taking stock of what is needed, what one has to work with and how well one has done) and *planning* (deciding how to use what one has).

Quite clearly, in many types of communicative activity, it would be difficult to identify any type of 'language' that could be said to give evidence of strategic ability, and in such cases, its removal by Bachman and Palmer to a position external to language knowledge would seem the most feasible way to deal with it. In most tests of reading or writing, for example, this ability may leave no tangible traces.

However, speaking is different. The processes – from intention to articulation – are going on in 'real time' (Levelt 1989), leaving abundant evidence of this fact, e.g. in the form of self-monitoring devices, appealers for help and communication strategies to overcome 'gaps' (Bialystok 1990). The particular demands placed on this ability by the medium of speaking are considered in more depth in 'Speaking' opposite.

The four-part model of CLA, developed for the EVA speaking testing situation, can be summarised as follows:

- MICROLINGUISTIC ABILITY
  the ability to access and use with some degree of correctness the essential systems of language at the level of the sentence/utterance and below, i.e. vocabulary, morphology, syntax and phonology
- TEXTUAL ABILITY
  the ability to make a text 'coherent with respect to itself', involving cohesion as the expression of semantic relations and the use of markers and routines that build structure into conversation as well as the organisational ability to structure information
- PRAGMATIC ABILITY
  the ability to use and interpret language in the way that it is typically used and interpreted by the society and in the particular situation in which the communication is taking place
- STRATEGIC ABILITY
  the ability to use devices to keep conversation going in face of difficulty and to check for, acknowledge and tackle potential problems in communication.

Describing the domain of CLA

So far, the discussion has referred to CLA as it may apply to any group of learners. Moreover, apart from certain specific references to the spoken language, notably in the case of strategic ability, the discussion has not been restricted to any one medium (i.e. spoken or written) of language; nor has it differentiated between productive and receptive language activity. However, in approaching the stage of operationalising language ability, i.e. describing its realisation in terms of actual behaviour, it is necessary to consider the fact that, firstly, what is being tested is speaking ability (in spoken interaction) and, secondly, the testees are 14–15-year-old Norwegian students. Both of these attributes determine the area or *domain* of CLA to be tested. This domain will be described in this section by considering first 'speaking' and then 'the situation of the testees'.

Speaking

Integration of language activities – active and passive, written and spoken – is increasingly acknowledged as pedagogically desirable (e.g. by Seda and Abrahamson 1990). Indeed, this integration reflects real-life communication. Spoken interaction always involves listening, and so the ability to interpret speech is as necessary as the ability to produce it. Moreover, the use of recent communication methods, such as e-mail, is fuzzying the boundary between spoken and written language.

However, while the formal differences between speaking and writing, such as those listed by Brown and Yule (1983b), may have become somewhat dated, it is still a fact that speaking is different from writing. As I sit at my computer, I can spend time on my e-mail message, check and reformulate it and ponder over my turn of phrase. What is more, (so far) my recipient cannot interrupt me, or help me out. It is these two fundamental, prevailing differences between speaking and writing – that speaking normally goes on in 'real time' as we decide what to say, and takes place together with someone else – that Bygate (1987) takes as a basis for identifying a set of skills specific to speaking.

Using Bygate's explanation as a starting point, this section will outline the particular skills needed to communicate successfully in speaking. The term 'skill' is used here, as it is by Bygate, to refer to the particular things we need to be able to do to function as speakers, and can be regarded as being superimposed, by the medium of speaking, on the components of competence outlined so far. When the components are operationalised, these skills will be taken into account.

Bygate calls the two sets of conditions of speaking which differentiate it from writing *processing conditions* and *reciprocity conditions* (1987: 7).

By processing conditions, Bygate is referring mainly to the time-pressure factor involved in speaking, and the fact that we have no record to consult, either as speakers or listeners. These two factors put the speaker at a seeming disadvantage compared to the writer. However, Bygate points out that this is compensated for through the other set of conditions – the fact that speaking is a reciprocal activity. Our interlocutor(s) can keep us right and help to prevent or clear up misunderstandings as they arise, which our readers cannot do. It is in order to cope with the first set of conditions, and to make use of the second set, that Bygate believes two distinct sets of skills are required, which he refers to as *production skills* and *interaction skills*, respectively.

Production skills can be regarded as the skills needed to help ourselves in speech production. As this is done under time-pressure, we need certain devices to facilitate our production, e.g. by playing for time, using verbal (or non-verbal) fillers or formulaic chunks of speech. Moreover, we need ways of compensating for the fact that things do not always run smoothly, e.g. by repetition or carrying out self-repair.

Interaction skills are those that take advantage of, or allow for, the fact that an interlocutor is involved. The first group of interaction skills involves knowing and being able to use *routines*, or predictable patterns. These routines may be those that help us to structure what we are saying so that the other person(s) will understand us better, e.g. by our chronologically ordering events in a narrative. On the other hand, routines may directly involve the other speaker(s), either in recurring exchanges, such as opening and closing a chance meeting or a telephone conversation, or in specific 'transactional' exchanges, such as making an appointment or buying clothes. The second type of interaction skills are *negotiation* skills. These skills are subdivided into those that negotiate meaning and those that manage the interaction itself. Negotiating meaning involves, for example, choosing the level of explicitness, or signalling understanding and acknowledgement of what the other speaker has said, or indicating a need for clarification. Managing the interaction can involve the agenda, e.g. introducing new topics or referring back to old ones, or the turn-system, by which we signal that we want to give, take or hold the turn.

Bygate's description of skills should perhaps be supplemented by the ability to use what Channell (1994) has called 'vague language', such as *sort of* or *I think*. Channell maintains that:

> research suggests that 'vagueness is present in a great deal of language use, and that therefore a complete theory of language use must have vagueness as an integral component.' (1994: 5)

Channell believes that we need vague language for two reasons. The first is that we may simply not be able to say exactly what we mean with the language

resources we have and in the time we have, and so need a way to signal that what we are saying must not be taken too literally. The other reason is that we do not always perceive the world in precise terms, so that we need a way to signal the 'fuzziness' of what we have in mind. This need, to show that the language we are using does not entirely 'match' what we have in mind, frequently arises and may be motivated either by a genuine inability to express exactness, or by the fact that we may want to avoid sounding pedantic or superior in some way.

This last point is to do with preservation of our interlocutor's (or even our own) 'face'. Brown and Levinson (1987) sum up this central notion of face as involving two basic universal desires: the desire to be unimpeded in one's actions (negative face) and the desire (in some respects) to be approved of (positive face) (1987: 13). What we say is frequently prompted by a wish to make the other person feel 'good', or at least to avoid making them seem insignificant, wrong, mean, stupid, etc. In fact, speaking is often carried out simply 'to be social' (Stenström 1994: 126). Empathising, showing interest, being polite or just 'keeping the channel open' are other ways of preserving face, and these social skills seem to warrant a place in the category of interaction skills.

Through the adaptation of Bygate's account to include the ability to signal vagueness and generally to be social, the following summary can be given of the skills that are specific to, or strongly associated with, speaking, bearing in mind that these skills equally involve production and interpretation:

- *skills required to 'play for time'*
  e.g. by using fillers or formulaic language
- *skills required to involve or acknowledge the interlocutor, or his/her utterances*
  e.g. by responding to the meaning of what s/he has said, establishing turn-taking roles, or by being polite or empathising in the interest of 'face'
- *skills required to structure, or 'place', utterances in the discourse*
  e.g. by using recurring or 'transactional' routines, or information-structuring conventions, or in linking what is being said to other parts of the discourse
- *skills required to cope with potential problems in communication*
  e.g. by choosing a level of explicitness, by checking for understanding or signalling misunderstanding, or by self-repair
- *skills required to express vagueness and lack of total commitment*
  through use of vague language such as *I think* and *sort of*, motivated either by genuine 'need' or in the interest of 'face'.

Little attention has been paid so far to the actual body of language required to put these skills into practice, although 'verbal fillers', 'formulaic expressions',

and 'vague language' – made up of certain words and expressions – have been referred to. 'Smallwords' were provisionally defined in Chapter 1 pages 1 – 6, as *small words and phrases, occurring with high frequency in the spoken language, that help to keep our speech flowing, yet do not contribute essentially to the message itself*. The explicit role played by smallwords in putting these skills into practice is returned to in Chapter 6 pages 122 – 156. However, even at this stage, it is evident that smallwords are key players in each of the skills outlined here, which are concerned with 'flow' rather than 'message'. Understanding, acknowledging, empathising, turn yielding, opening and closing, locating the relevance of an utterance in the discourse and carrying out self-repairs are just some of the things we signal with smallwords when we speak. In operationalising spoken CLA, not only should skills specific to speaking be considered but also the language needed to put these skills into effect, and thus smallwords must be given attention in the operationalisation process.

Before leaving this section, it is appropriate to comment on the relationship between the skills identified here and the components of ability worked out in the previous section. No one-to-one relationship is possible or desirable, but, clearly, no component of CLA, as it is operationalised in 'Operationalising components of CLA' pages 49 – 55, should be untouched by what has been outlined here. The textual component should include, for example, some ability to use routines, structure one's own turn, make links within the discourse and signal turn management. The pragmatic component should include some ability to select routines appropriately and to know how to be 'social' in conversation and which of (and when) the different signals used in conversation, including those of vagueness, are appropriate. The strategic component should include some ability to use the skills necessary to cope with potential breakdowns. And the microlinguistic component should include having access to the particular words, expressions and structures necessary to put all these skills into practice.

The situation of the testees

The situation of the testees affects several aspects of the domain of CLA, notably the *topics* they are expected to be able to address, the *conditions* under which they should be able to communicate and the *functions* they can be expected to perform through their speaking, as well as the *level* of CLA that can be expected of them. In the case of students taking the EVA speaking test, these factors are defined partly from the school curriculum, backed up with reference to other literature, partly from knowledge of the kind of spoken English communication students are likely to take part in in their near future, and partly from experience of students' performance. On the basis of this, the following framework has been drawn up:

TOPICS

The topics in this framework are based principally on what is laid down in M87 the Norwegian national school curriculum (valid at the time of test development and trialling), which coincides to a very large extent with the 'specific notions' outlined in van Ek's *Threshold level: 1990*:

- personal identification
- school life and 'classroomspeak'
- home and local environment
- food and drink
- free time – hobbies, sport and culture, social life
- shopping and services – eating out, etc.
- travelling – managing abroad and entertaining/helping foreigners in Norway
- education and work
- personal relationships.

CONDITIONS FOR SPEECH

The conditions that affect the way we speak consist of four variables, according to Weir (1993), and will be adhered to in this study:

- purpose
- interlocutor
- setting
- channel.

(1993: 37 – 38)

In the case of the EVA testees, the overarching purpose and setting of their speech situations are defined by the fact that the students are being tested in school. However, tasks are designed so that they give an 'internal' purpose, and in some cases a setting is simulated. These purposes and settings are associated with the topics being discussed. Interlocutors can be differentiated according to age group, relative status, and degree of familiarity. Students should be able to talk to youngsters and adults, people with equivalent or superordinate status, and both friends/family and strangers. The most normal channel is face-to-face, but students are expected to be able to cope with the telephone.

LANGUAGE FUNCTIONS

While a piece of communication will normally have an overall purpose, e.g. being friendly over lunch, or inviting someone to a party, each of the actual speech acts produced, as the talk proceeds, performs a function, e.g. prefacing, asking or warning. These functions are too numerous to list here, and so will be considered as belonging to categories, or 'macrofunctions'.

The macrofunctions defined as relevant for the EVA students are taken from van Ek and Trim's (1993) six categories of 'what people do by means of language':

- imparting and seeking information
- expressing and finding out attitudes
- getting things done
- being social
- structuring discourse
- communication check-and-repair.

(slightly adapted from 1993: 22)

The reason for selecting this particular model of macrofunctions is based largely on its match with what is exemplified in the school curriculum, M87, and what has been discussed as pertinent to speaking on pages 43 – 46. Moreover, it is consistent with much of what has been described as making up the major functions of language in communication.

The first three of these macrofunctions seem to cover more or less what Halliday (1994: 106) defines as three major processes that go into our building, through language, a mental picture of 'reality': broadly, *being, sensing* and *doing*. A small adaptation made to van Ek and Trim's model is the dropping of the term 'factual' before 'information' in the description of the first macrofunction. This has been done in order to accommodate the expression of imaginative ideas, in line with impending adaptations to the school curriculum and with what Bachman calls the 'imaginative function of language' (1990: 94). The macrofunction of *getting things done* is given some coverage in M87 under the heading 'getting others to do something' (inviting, asking for help, offering, warning and advising) and seems to correspond to what Bachman terms 'manipulative functions' (1990: 93), i.e. those 'in which the primary purpose is to affect the world around us'.

The fourth macrofunction, here called *being social* (as opposed to van Ek and Trim's 'socialising'), as was suggested in the previous section, can be a significant end-in-itself in speaking, and this macrofunction can be regarded as a component of what has been called the 'interpersonal' metafunction of language (Halliday 1994).

The two remaining macrofunctions – *structuring discourse* and *communication check-and-repair* (c.f. van Ek and Trim's 'communication repair') might be regarded, to a large extent, as the realisations of (respectively) textual and strategic ability in speaking, and have already been given attention in the previous section on speaking skills.

The six-part framework of macrofunctions should be applied to the final description of CLA to the extent that all the macrofunctions should be performable by students. The range of the actual (micro-)functions cannot

be defined precisely but can be expected to be made up of those functions outlined and exemplified in M87, together with a range of context-specific functions, such as buying, booking, ordering, and those related to the forming of personal relationships. These can, however, be supplemented by any functions that a student can reasonably be expected to need in order to cope with the topics drawn up.

LEVEL OF ABILITY

The students taking the EVA speaking test are typically 14 or 15 years old, and have been learning English for between four and five years, with two lessons per week. The students have generally been subjected to a high level of media exposure to spoken English. At this stage, it is necessary to be cautious about describing level of ability; the outcome of the test development and the actual testing will shed more light on this. However it is possible to define a level of ability which students are expected to aspire to, followed at a later stage by the building of band-scale descriptors for various levels of ability.

The level of ability aspired to can be broadly defined as follows. Students should be able to take the initiative and keep themselves going with minimal prompting, show a varied, independent and idiomatic language use, with some knowledge of basic social and situational conventions. They should be able to be reasonably polite and show some sensitivity to others in their language use, and should show awareness of the more common conversational or transactional routines.

To sum up, by reference to literature and to the school curriculum, supplemented with experience with the students and their potential 'likely' English-speaking situations, this section has outlined a set of features that constrain the domain of CLA in terms of:

- topics of speech
- conditions under which speaking occurs
- functions typically performed by speakers
- level of ability (aspired to).

Operationalising components of CLA

In this section, the aim is to produce a set of operationalised components of CLA, which describe the actual behaviour that indicates that a student 'has' each of the component abilities that are theoretically defined as making up CLA. Only by having a definition of this behaviour is it possible to draw up specifications for the test tasks for eliciting it, as well as for the rating instruments which have the job of describing the behaviour at different levels of ability.

Four components of CLA have already been defined in 'A suitable model of CLA' pages 39 – 42. In this section, each component will be taken in turn and discussed in the light of what has been described as the domain of CLA in the previous section. The discussion will be summarised as a description of the way the components can be operationalised, i.e. expected to be manifested in the language of students.

It must again be emphasised that what is described is based on the ability that students aspire to according to the school curriculum, and to what actual triallings and consultation with teachers have revealed to be the 'top' level normally reached by students (i.e. not counting bilingual students). Most students will only demonstrate this ability in part. It must also be stressed that the components are not to be regarded as watertight compartments in any way. There is very real interdependence between them, as should become obvious as the descriptions proceed.

Operationalising microlinguistic ability

On pages 39 – 42, the component of *microlinguistic ability* was defined as:

> the ability to access and use with some degree of correctness the essential systems of language at the level of the sentence/utterance and below, i.e. vocabulary, morphology, syntax and phonology.

The terms 'access' and 'use with some degree of correctness' have been chosen deliberately here to suggest that this ability is to do with 'having' a stock of language and using it without making too many mistakes. The greater the stock used and the fewer mistakes made, the higher will microlinguistic ability be rated.

As to the question of how mistakes should be tolerated, it has to be borne in mind that, as was pointed out on pages 43 – 46, speaking is processed 'on-line', putting the speaker at a disadvantage (relative to the writer) in terms of consistently being able to produce 'correct' forms. This point is reinforced by Sharwood Smith (1994), who identifies as 'control variability' the language inconsistencies 'caused by factors having to do with the on-line processing of competence' (1994: 109). In other words, the learner can 'know' correct forms with varying degree of control over, or accessibility to this knowledge. In times of stress, distraction or any other kind of 'processing overload' the learner may not be able to perform in the way they 'know' they ought to. In the case of morphology, for instance, Pienemann and Johnston (1987) say of the third person *-s* marker: 'students [...] can frequently be induced to produce it in the classroom, yet the moment they begin speaking spontaneously, except in the case of all but the more advanced learners, it disappears' (1987: 81). It must be concluded that, in terms of accuracy, microlinguistic ability makes fewer demands of the speaker than of the writer.

As regards 'which' language' students should have access to, this should include the stock of words and expressions that specifically enable students to put the skills outlined in 'Speaking' pages 43 – 49, into practice. This consists largely of smallwords, which include 'vague language', as well as common formulaic expressions that facilitate speech.

In order to function in their own particular situation, these testees should, of course, also have the basic and specialist vocabulary and structures to talk about and 'operate within' the topics and associated situations as outlined in 'The situation of the testees' pages 46 – 49. They should also be able to perform functions representing all the macrofunctions in at least a straightforward, or 'transparent' way.

Moreover, in order to communicate successfully, the sounds and intonation patterns as well as other prosodic features must be mastered to the extent that they allow the 'message' to come across fully.

Operationalising textual ability

Textual ability was defined in 'A suitable model of CLA' pages 39 – 42, as:

> the ability to make a text 'coherent with respect to itself', involving cohesion as the expression of semantic relations and the use of markers and routines that build structure into conversation as well as the organisational ability to structure information.

This definition implies that the speaker must be able to bring about coherence both in 'long turns', where turn-internal cohesion, e.g. by ordering and linking ideas, is involved, and in 'short turns', where across-turn cohesion, e.g. by signalling acknowledgement or turn-taking, is involved. Brown and Yule (1983a) voice concern that 'it is currently fashionable in language teaching to pay particular attention to the forms and function of short turns', maintaining that 'simply training the student to produce short turns will not automatically yield a student who can produce long turns' (1983a: 19 – 20).

The ability to create both turn-internal and across-turn cohesion is reflected in the speaking skills outlined in 'Speaking' pages 43 – 46. Turn-internal cohesion may be brought about through the use of information-structuring conventions, and linking devices, such as pronouns and deictics, as well as those smallwords referred to by Stenström (1994) as *discourse markers*, which 'are used to organise and hold the turn and to mark boundaries in the discourse' (1994: 63). In the case of the EVA testees, suitable text types that would illustrate the ability to create turn-internal cohesion could include short narratives and descriptions.

Across-turn cohesion is contributed to by the use of recurring routines, such as those exemplified by Stenström (1994) in her description of 'exchange procedures' (1994: 84 – 126). Devices to involve the interlocutor or his/her

utterances also play a role in this type of cohesion, and these largely consist of the smallwords referred to by Stenström (1994) as *interactional signals*, which 'are used to start, carry on and terminate the conversation', e.g. by appealing for and giving feedback, giving response, and involving the listener in the conversation (1994: 61). Suitable tests for EVA testees that exemplify the ability to create turn-internal cohesion could include giving and receiving instructions, and taking part in discussions.

Thus, students should be able to create cohesive, structured speech in 'holding the floor' and should be able to use interactional signals such as *right* and *you know,* which 'play a crucial role for a smooth interaction' (Stenström 1994: 61). Achieving some degree of smoothness, in fact, seems to be implied by textual cohesiveness, and students should be able to speak at a rate, and with a degree of connectedness, that allows what they say to 'hang together' in the conversation and allows communication to proceed unhampered. This smoothness can be assisted by the use of formulaic expressions (Bygate 1987; Pawley and Syder 1983), and smallwords such as *you know*, acting as a verbal fillers when 'speakers need more time' (Stenström 1994: 69), as well as the use of vague language (Channell 1994).

Operationalising pragmatic ability

Pragmatic ability was defined in 'A suitable model of CLA' pages 39 –42, as:

> the ability to use and interpret language in the way that it is typically used and interpreted by the society and in the particular situation in which the communication is taking place.

Using language in a way typical of society implies firstly that students should be able to perform functions of the macro-types identified in 'The situation of the testees' pages 46 – 49, in the way they are normally carried out, i.e. 'idiomatically'. Moreover, as one of the most important macrofunctions in speaking is *being social*, students should be able to use their language in a way that gives regard to their interlocutors' 'face' in the conventional ways. As was shown in 'Speaking' pages 43 – 46, this involves speakers' expressing themselves suitably politely, and by using what Stenström (1994) calls 'empathisers', such as *you see*, 'inviting the current listener to take an active part, as it were' (1994: 127). And being social can also involve the use of vague language, for instance in giving opinions, when the student should be able to 'hedge' or soften the force of what s/he is saying, e.g. with *I think* , to 'avoid going straight to the point, avoid appearing authoritative, and avoid committing him/herself' (Stenström 1994: 128).

Using language appropriate to the particular situation in which the communication is taking place implies being able to adapt to the conditions affecting speaking, identified on pages 46 – 49, as purpose, interlocutor,

setting, and channel. The purpose and setting, in the case of the EVA testees, can be expected to vary widely, but can be regarded as being associated with the topics described, and bound by what are feasible purposes and settings for students of this age, when using English. This involves the ability to use transactional routines linked, for example, to travel or shopping. Furthermore, the students can be expected to be able to use language appropriate to interlocutors of their own age and adults, both familiar and unfamiliar. They are also expected to be able to communicate by telephone as well as face to face.

Operationalising strategic ability

Strategic ability was defined in 'A suitable model of CLA' pages 39 – 42, as:

> the ability to use devices to keep conversation going in face of difficulty and to check for, explain and tackle potential problems in communication.

Strategic ability, like textual ability is to do with 'smoothness' in communication. But while textual ability is regarded as involving smoothness that is brought about by producing coherent and connected speech, strategic ability is regarded as being concerned with smoothness when it is threatened by breakdowns between form and meaning. Four types of such breakdowns can be identified. The first occurs when the speaker is unsure which form to use to express a meaning, and this brings into play what have conventionally become known as *communication strategies*. The second can be thought of as the breakdown between what the speaker has just said and what s/he meant to say, the third between what the speaker has meant and how his/her interlocutor has interpreted the message, and the fourth between what the interlocutor has just meant and how the current speaker has interpreted it.

All of these (actual or potential) situations call on what have been described in 'Speaking' pages 43 – 46, as *skills required to cope with potential problems in communication*, and putting into effect these skills can involve what has been cited in 'The situation of the testees' pages 46 – 49, as the macrofunction *check-and-repair*. In the event of imminently threatened/looming breakdowns, students are expected to be able to use the common forms of signalling and coping with these. In 'self-repair', students should know how to signal that they are carrying this out, e.g. with *I mean*. In checking that their interlocutors have understood what they mean, students should be able to do this, e.g. through the use of *all right?* And in the event of not quite understanding their interlocutors, students should be able to ask for repetition and clarification.

However, students should also be able to avoid breakdowns through the use of 'communication strategies', which, according to Bialystok (1990), are triggered by 'gaps' in our knowledge, and 'can take many forms – a word,

a structure, a phrase, a tense marker, an idiom' (1990: 1). These strategies have been given a considerable amount of attention in the past decade or so, e.g. by Færch and Kasper' (1983), Kellerman, Bongaerts and Poulisse (1987), Tarone and Yule (1989) and Bialystok (1990), who have manipulated them into a range of taxonomies. For the purposes of this study, however, where ability is in focus, strategies need to be classified only according to what speakers do (rather than what they say), and this is the basis for Bialystok's (1990) classification into *analysis-based* strategies and *control-based* strategies.

Analysis-based strategies are put into effect when the communicator (who could, in fact, be a native speaker), finding him or herself unable to provide the 'label' for a concept, turns to the process of analysis of the concept itself, e.g. *robin*, producing for instance *a bird with a red breast.* Our attention is on the analysed meaning of the message. Control-based strategies, on the other hand, keep the concept intact, but adapt the means of expression, e.g. using a foreign word or phrase, miming or pointing. In other words, we attend primarily to selecting a form of expression to convey our message.

In the case of the EVA testees, while it would be wrong to discourage control-based strategies using, for instance, body language, those involving resorting to Norwegian words are clearly unacceptable. On the other hand, students are expected and encouraged to be able to use analysis-based strategies.

Some conclusions on CLA and the significance of smallwords

There is clearly great interdependence between the components of CLA as described here. None of them would get us far in communication without the others. Microlinguistic ability basically involves students having a certain stock of language knowledge at their disposal, so that they are able to produce the bits of message necessary for communication to come across. Textual ability involves putting the bits together so that the communication functions smoothly as a 'whole'. Pragmatic ability involves selecting the most appropriate bits of language to use, so that the communication functions 'normally'. And strategic ability involves coping when the communication is threatened with breakdown.

As far as the 'body of language' necessary to put these abilities into use is concerned, certain generalisations can be put forward on the basis of the discussion in this chapter. This language body can be viewed as being made up of 'message' and 'non-message' language. The former is the language needed to convey the actual content of the message. Microlinguistic ability depends on a knowledge of this language, textual ability depends in part on

knowing how to order and connect it, and pragmatic ability depends largely on knowing how to make appropriate selections from it.

Otherwise, however, the abilities, as they have been described here, involve the knowledge and use of a body of non-message language, i.e. things we say that do not directly contribute to the message itself. As is apparent from the discussion here, this body is largely made up of what have been presented as 'smallwords'. Textual ability draws on smallwords to help play for time and keep going, as well as to structure both long- and short-turn conversation. Pragmatic ability needs them to help give regard to the other speaker's face, e.g. through hedging. And strategic ability uses them to signal the various types of check-and-repair work that keep communication running smoothly.

A closer analysis of the role of smallwords in communication will be left to Part Two. However, already at this stage it is apparent that no study of how spoken CLA should be elicited and assessed can afford to neglect a body of language so fundamental to spoken communication.

Summary

This chapter has recounted the process carried out at a stage prior to the EVA test development, in which components of CLA were operationalised. The process began by working out a definition of what is normally considered to make up communicative competence. Next, the domain of language use relevant to the EVA speaking test was defined, by considering the particular skills demanded by the medium of speaking, and the particular situation of the EVA testees. The chapter culminated in putting the findings of these deliberations together and producing a set of four operationalised components of CLA, expressed in terms of the language behaviour that can be expected of EVA students who 'have' each component of ability.

No attempt has been made to group components according to what might be jointly measurable, in that a score on one component (or part of it) might predict the score on another. Nor has any speculation been made as to whether some components or behaviours might be developed earlier than others. These issues relate to the validity of the band-scale descriptors and will be touched on in 'The validation process' pages 65 – 87.

What has been achieved, however, is that a description has been given of the kind of behaviour the EVA test should aim to elicit and base rating procedures on. Moreover, the description is comprehensive in that it takes account of all aspects of communicative competence as it has been typically described in recent literature, but it is compact enough to provide a usable blueprint for the test specifications. Furthermore, it provides a checklist for the validation process, whereby the question 'does the test test what it is supposed to test?' will be systematically addressed. Now, at least, we know what it is that the EVA speaking test is supposed to be testing.

To sum up, components of CLA have been operationalised in this section as follows:

MICROLINGUISTIC ABILITY implies being able to spontaneously:

- access and use with some accuracy the stock of words and expressions necessary to put the skills specific to speaking into practice – these largely being common smallwords and formulaic expressions
- access and use with some accuracy the stock of general and specialised vocabulary and language structures to talk about and 'operate within' the specified topics and associated situations
- perform functions of all the six specified macro-types, in a straightforward, transparent way
- produce sounds and intonation patterns well enough to allow the message to come across in full.

TEXTUAL ABILITY implies being able to spontaneously:

- produce turn-internal cohesion in such texts as descriptions and narratives, by ordering information conventionally, and by using links, such as pronouns and deictics (e.g. *over there*) and organising devices such as smallwords acting as discourse markers (e.g. *well* and *right*)
- produce across-turn cohesion in such texts as instructions and discussions, by the use of common conversational routines, and through smallwords acting as interactional signals (e.g. *okay* and *right)*
- speak smoothly, i.e. at a rate that is not detrimental to communication and without excessive hesitation. This smoothness may be assisted by formulaic expressions, verbal fillers (e.g. *you know*) and vague language (e.g. *sort of*).

PRAGMATIC ABILITY implies being able to spontaneously:

- perform functions of all the six specified macro-types, in a conventional, 'idiomatic' way
- show regard to interlocutor's face, e.g. through the conventional use of empathisers, politeness expressions and hedges (e.g. *a bit*)
- use transactional routines according to the purpose and setting of the speaking situation
- adapt language according to the age and familiarity of the interlocutor
- communicate by telephone as well as face to face.

STRATEGIC ABILITY implies being able to spontaneously:

- use communication strategies that primarily employ English (i.e. analysis-based) only resorting to other (control-based) strategies as long as these do not involve using non-English forms
- carry out self-repair, check understanding on the part of the interlocutor, and indicate own lack of understanding, using the (small)words (e.g. *I mean* and *you know*) and other expressions normally employed to carry this out.

4 Validation of the test 'as it stands'

In this chapter the a priori validation proper of the EVA speaking test is carried out, i.e. the test is inspected with respect to the way it seems to be valid as it stands when going out to schools. The six-part framework worked out in 'A unified framework for validation' pages 28 – 31, based on Messick's (1996) central aspects, is used to provide a structure for the validation process. Within each of the six aspects, all the potential sources of invalidity are considered with respect to the extent to which they have been safeguarded against in the test development and whether this safeguarding is supported either in the literature or by primary empirical evidence, and the extent that further, a posteriori, validation or amendment to the test seems appropriate. The relative strengths and weaknesses of these six aspects of validity in the test are summed up, with recommendations for further investigation in both the present and future studies.

As it is impossible to launch a thorough investigation of every aspect of test validity on the basis of a set of test materials used in a single round of actual testing, the intended outcome of this investigation will be restricted to:

- judgements on the strengths and weaknesses in the test's validity which it is possible to make at this stage on the basis of the test as it stands (a priori)
- recommendations for investigation able to be carried out in the present study using scoring data and transcripts of students' performance (a posteriori)
- recommendations for future investigation, beyond the scope of the present study, e.g. using data from revised test versions, test transcripts or surveys among test users.

Fundamental to the validation is the model of CLA operationalised in Chapter 3 pages 33 – 57. This provides a detailed description of the 'thing' – language ability – that is assumed to be being tested; without referral to this, the validation process would be unable to provide an answer to the pivotal question 'does the test test what it is supposed to test?'. The extent to which this model of CLA appears to be reflected in the test tasks and scoring instruments is revealed in the course of the validation process, and is also a recurring theme in the conclusion to the chapter.

The aims and purposes of the EVA testing

It was stated in 'Validation – an overview' pages 10 – 28, that 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). Chapter 3, pages 33 – 57, took up the first question, but before the validation process can be embarked upon, it is necessary to present the aims and purposes of the EVA speaking test, which were briefly introduced in 'The test' on page 3, followed by an outline of its specifications.

The main aim of the EVA testing, laid down in 1993, was to give a diagnostic profile of the strengths and weaknesses of students' communicative language ability, primarily at the level of the four macroskills of listening, speaking, reading and writing. Furthermore, a more detailed profile of ability was to be presented in the case of speaking and writing. As the tests might be used to survey language ability, either in districts or nationally, a further aim was to translate scores on all macroskills onto a numerical grading system. In the case of the speaking test this meant that the test feedback would consist of both a profile and a grade, based on band scales, for each student.

The principal purpose was to assist teachers in identifying students' needs, in terms of where they might require extra help or tasks at a more advanced level. A secondary purpose was to enhance the competence of teachers and students in assessment itself, through the introduction of innovative procedures and methods into the language classroom. An ultimate purpose was to enhance the learning process, through increased awareness among students of what they could do, and should aspire to do, and through a more informed approach on the part of their teachers.

Speaking test specifications

The EVA speaking test can be thought of as being made up of two main parts – *elicitation procedures* (to do with the test tasks) and *scoring procedures* (to do with feedback on performance). In this section, the specifications for each of these parts are discussed and an outline of their formats is presented. The operationalisation of CLA for the student group being tested (see 'Operationalising components of CLA' pages 49 – 55,) has been fundamental to the development of both procedures. However, the specifications were also influenced by what was laid down in the testing mandate, and were worked out in close consultation with a team of researchers/teacher-trainers in the West of Norway, with several rounds of experimenting and trialling in local schools.

Specifications for elicitation procedures

The fundamental requirement of the language elicited in the EVA speaking test is that it should contain evidence of the extent to which testees have all the components of CLA, as they are operationalised in 'Operationalising components of CLA' pages 49 – 55,. In order to demonstrate microlinguistic ability, therefore, students must be given the chance to converse on topics randomly picked from those listed in 'The situation of the testees' pages 46 – 49, and to perform the macrofunctions listed in that section. They should also be given the opportunity to show whether they can pronounce more difficult words and to use intonation which supports the message; this may be effected by the inclusion of a short text to be read aloud. To show their textual ability, students must be allowed to participate in short conversational exchanges as well as to hold the floor, taking longer turns which require internal structuring, such as in describing and narrating. To demonstrate their pragmatic ability, students should be given the chance to use devices that take care of 'face', e.g. hedging to soften the force of opinions. They should also be placed in situations, e.g. through role-play, where they can show their ability to use conventional transactional routines, such as in buying, as well as their ability to adapt to the range of conditions outlined on pages 46 – 49, such as when speaking to an unfamiliar adult, or using a telephone. In order to demonstrate strategic ability, even the best students should be put into situations where communicative problems are likely to arise. This can be brought about by deliberately introducing problematic elements into the dialogue, e.g. through the use of pictures, designed so that it can be assumed that at least one of the speakers is unlikely to have full access to the appropriate vocabulary.

With the above requirements in mind, the following test format has been worked out. The test is made up of three tasks, and takes 20–25 minutes. It is carried out with students in pairs. A third person, usually an adult, is present as director. The director has a 'script' for the actual direction (Appendix E pages 282 – 285) and written instructions on how to proceed (Appendix F on page 286). Each student has a test booklet (Appendix A pages 268 – 273) containing instructions and all necessary props and information for the tasks. The students' written instructions serve mainly as additional support, as the director gives all the necessary input (e.g. instructions and prompts for discussion). Each task contains specific sub-tasks for each of the partners. A practice test is issued, as is a video recording of performances.

The director (normally either the teacher or a resourceful third student) is responsible for practical problems, such as recording and guiding the students through the task, using the provided script, and stepping in where necessary to keep the test moving. The script (Appendix E) ensures that all students taking the test receive very similar input, and that each student is given a chance to carry out his/her side of the tasks irrespective of the other student's

performance. Ideally, the test is carried out in a small room away from the rest of the class, with the two students placed together, half-facing each other and both able to see the director. The performance is recorded. Appendix A pages 268 – 273, shows a student's task booklet for one of the three versions of the test. The task design is as follows:

Task 1
In this task, both students have a series of pictures, which together make up a narrative. Each student has to describe a scene, tell his/her half of the story and comment on how the characters felt at different points in the narrative. The students are given questions on the connection between the two halves of the narrative. A discussion follows, in which the students are each given the opportunity to express opinions and to agree or disagree with each other. This discussion is based on a personalised aspect of the topic presented in the pictures.

Task 2
This task involves each student's giving the other a set of instructions, based on pictures. The task is designed so that fairly inaccessible vocabulary is called on from time to time. The student receiving the instruction is invited to ask questions to clear up problems, and to follow up the task by recapping the main points.

Task 3
This task is conducted as a 'semi-role-play' and has two parts. For each part, the situation is described in a short, written text, and the dialogue is presented with one role written out and the other given as a series of functions to be performed. These functions include 'socialising', e.g. introductions, thanking and finishing off the conversation, and 'getting things done', e.g. inviting or requesting help. Students are given time to look over and think about the task, before performing it. As two sub-tasks are provided, each student is able to take a turn in reading aloud the introductory text and the fixed speech, and in improvising the missing dialogue. One of the two conversations is done by telephone and one is a transaction, such as buying something or applying for a holiday job. The student using fixed (read) text normally plays the role of an adult. However, in the other, improvising role, the student is never called upon to be anyone but him/herself.

This combination of tasks has been designed to ensure that any version of the test will elicit evidence of all components of CLA, covering the full range of conditions of speech, and requiring the performing of all the macrofunctions identified in 'The situation of the testees' pages 46 – 49. The test versions, of which there are three so far, should ensure a wide coverage of the topics listed on pages 46 – 49. The way topics, conditions, macrofunctions and components of CLA are specifically intended to be covered by the tasks

is illustrated in Table 4.1 (although, of course, other functions will be performed, and components of ability activated in the course of tasks):

Table 4.1: Coverage of topics, conditions, macrofunctions and components of CLA in the design of tasks

	task 1	**task 2**	**task 3**
topic	free selection	free selection	free selection
conditions			
purpose	describe, narrate, discuss	instruct/learn from instructions	free selection (manipulated)
interlocutor	familiar – peer/adult	familiar – peer/adult	familiar and stranger – adult and peer (manipulated)
setting	school/test setting	school/test setting	free selection (manipulated)
channel	face-to-face speech	face-to-face speech	face-to-face speech (read aloud) speech telephone (manipulated)
functions highlighted	Imparting and seeking information. Expressing and finding out attitudes. Structuring discourse.	Imparting and seeking information. Communication check-and-repair. Getting things done.	Being social. Imparting and seeking information. Getting things done.
component of CLA highlighted	textual microlinguistic (pragmatic)	strategic microlinguistic (pragmatic)	pragmatic microlinguistic

Specifications for scoring procedures

The scoring procedure is shaped by two main forces, the first being the way CLA has been operationalised and the second being the purpose of the testing, i.e. what was laid down in the mandate from the Ministry of Education. This means that all components of CLA, as operationalised in 'Operationalising components of CLA' pages 49 – 55, must be taken into account in the scoring procedure. Moreover, the feedback must satisfy the test's purposes, as mentioned in 'The aims and purposes of EVA testing' page 59. This means that, firstly, detailed diagnostic information must be given on a student's strengths and weaknesses, which can lead to improved learning (by drawing attention to what goes into speaking, and which aspects particularly need to be worked on) and that, secondly, a numerical overall grade must be computable.

In order to satisfy the dual purpose of diagnosing and grading, two scoring instruments have been developed: a 'performance profile' (Appendix B pages 274 – 276) and a pair of band scales (Appendix C pages 277 – 279). Ideally two raters should be used in assessing each performance, but in many cases, for practical reasons, scoring will be carried out solely by the class teacher who will probably also have been the test director. A handbook is issued for teachers/raters, with guidelines for assessment (Appendix D pages 280 – 281), and a video recording shows students at different levels of ability. Filled-in performance profiles and band scale ratings are included for the six students on the video.

PERFORMANCE PROFILE

The performance profile for each student should be filled in while the rater listens to a recording of the performance, selecting a level in a series of 'tick the box' statements on individual tasks (chronologically following the progress of the test) and finally on the performance overall. The statements together are designed to compile a detailed profile of the performance, reflecting all the components of CLA as operationalised in 'Operationalising components of CLA' pages 49 – 55. Moreover, certain statements are included that refer to the actual 'execution', i.e. general task achievement and fluency, and no association is made here between these aspects of performance and components of CLA. The set of statements can be summarised as covering the following elements, which are grouped here principally according to the component of ability they seem most closely associated with, although, on the actual form, this grouping is not used:

statements associated with microlinguistic ability
- variety and independence in language choices
- correctness and adequateness of language structures and vocabulary
- pronunciation and intonation as supporters of message

statements associated with textual ability
- linking and coherence of ideas in describing and narrating

statements associated with pragmatic ability
- indication of appropriate politeness, friendliness and interest, through intonation and choice of expression
- observing cultural conventions, either linked to politeness (e.g. use of *please*) or to 'neutral' language, such as the reciting of a telephone number
- using conventional expressions associated with particular situations
- idiomaticity, appropriateness

statements associated with strategic ability
- non-dependence on Norwegian for overcoming language 'gaps' and other communication problems

statements associated with general task achievement and fluency

- putting over the message independently in description, narration, opinions, expression of feelings, giving instructions and performing the interaction involved in role play
- taking the initiative and making a significant contribution
- keeping going with a reasonable rate of flow.

The purpose of the performance profile is twofold. Firstly it presents a picture of the student's performance that reveals strengths and weaknesses in detail, and so can be used by both the teacher and the student to ultimately improve the learning situation. Secondly it can be used by the rater/teacher to help decide where to place a student on the band scales.

BAND SCALES

Through the use of the two band scales (*message and fluency* and *language structures and vocabulary*), the overall level of ability of each student can be judged. Each scale is divided into three broad bands of ability, i.e. below, just over or clearly above a 'minimal level' which is defined as adequate for getting the basic messages across. Each of these three bands is further subdivided into two levels yielding a total of six levels on both criteria scales. The performance of the student is matched as closely as possible to a level on each of the two criteria scales, yielding a rough general level, and pronunciation and intonation is then taken into account to arrive at a final overall level of performance. The rater is instructed to give *message and fluency* a greater weighting than *language structures and vocabulary* in the case of difficulty in deciding the final level. This decision has been made on the basis of a consensus 'gut feeling' voiced by teachers that, in the case of spoken communication, getting the message across reasonably smoothly is more important than being mistake-free or using sophisticated language. Moreover, it creates a balance within the EVA testing as a whole, since in the case of the writing test the reverse advice is given.

The band scales are made up of statements that are intentionally positive as far as possible. What is more, they reflect all the components of CLA to some extent, although in less detail than in the performance profile.

MESSAGE AND FLUENCY

Statements in the *message and fluency* band scale can be summarised as relating to the following elements:

- initiative and contribution
- managing to put across the essentials of the task
- flow and hesitance
- ability to keep going without help
- linking and cohesion

LANGUAGE STRUCTURES AND VOCABULARY

Statements in the *language structures and vocabulary* band scale can be summarised as relating to the following elements:

- accuracy, variety and independence in language choices
- adequacy of vocabulary
- idiomaticity in language choices
- appropriacy in style and degree of politeness
- degree of coping without recourse to Norwegian.

Pronunciation and intonation are considered as adjusters to the final grade.

The use of two scales, containing the above elements, was triggered by the need to take the focus of assessment away from language correctness and sophistication (in a culture where assessment of written language has dominated) and on to other aspects of communication, more pertinent to spoken language use. At the same time, it was felt necessary to limit the number of scales as far as possible. The traditional language-fluency divide, familiar to most teachers, seemed a sensible framework on which to build the scales. A possible association between the scales and individual components of CLA will be discussed in 'The structural aspect of validity' pages 74 – 82.

The band scales, besides yielding a grade, also give a 'more-or-less' picture of the ability of a student, and a means of seeing what should be done in order to pull his/her performance to the next level.

The validation process

The process of a priori validation is carried out within the framework described in detail in 'A validation framework' pages 30 – 31. This framework is based on Messick's (1996) six central aspects of construct validity, as defined in 'Six central aspects of validity' pages 29 – 30, and has the following six parts:

- the CONTENT aspect of validity
- the SUBSTANTIVE aspect of validity
- the STRUCTURAL aspect of validity
- the GENERALISABILITY aspect of validity
- the EXTERNAL aspect of validity
- the CONSEQUENTIAL aspect of validity.

By considering all the potential threats to each aspect of validity, as identified on pages 30 – 31, the validation process will be carried out aspect by aspect, through inspection of the test as it stands.

The CONTENT aspect of validity

In 'A validation framework' pages 30 – 31, it was maintained that the content aspect of validity may be threatened by:

- faulty or incomplete operationalisation of components of the model of CLA
- poor sampling of the language associated with the underlying theoretical model and domain of CLA, when making tasks
- unclear instructions or unfamiliarity of format (which prevent tasks from being done as intended)
- test methods and procedures that may prevent testees from performing in the way intended
- test bias associated with cultural background, background knowledge, native language, ethnicity, age, or gender.

FAULTY OR INCOMPLETE OPERATIONALISATION OF COMPONENTS OF THE MODEL OF CLA

The process of operationalising the model of CLA was given considerable attention in 'The situation of the testees' pages 46 – 49, where it was made apparent that the school curriculum was closely consulted and the situation of the student group being tested was taken into consideration in this process. Moreover recognised works of literature were drawn on, in both the field of speaking (e.g. Brown and Yule 1983b, Bygate 1987 and Stenström 1994) and in that of defining and testing communicative competence (e.g. Hymes 1972, Canale and Swain 1980 and Bachman 1990). Thus, there seems no reason to conclude that this operationalising process has been carried out in a faulty or incomplete way.

POOR SAMPLING OF THE LANGUAGE ASSOCIATED WITH THE UNDERLYING THEORETICAL MODEL AND DOMAIN OF CLA, WHEN MAKING TASKS

The content aspects of validity primarily concern the extent to which the sample of language elicited by the test tasks gives representative evidence of language ability as this is activated in the domain of TLU. This means that tasks are designed so that optimal responses should include all the kinds of language associated with the components of CLA as they are operationalised in 'Summary' pages 55 – 57. For example, a competent student will produce fairly accurate language that draws on a stock of words and expressions sufficient to talk about and operate within the topics randomly selected from those specified for the domain of CLA. The language will be appropriate to the varying situational contexts and will contain cohesive devices within and across turns. The language used in performing speech acts associated with any of the six macrofunctions will be fairly typical of the language of native

speakers when performing these speech acts. The language of a less competent student will contain less evidence of the student 'having' CLA, but this will be a characteristic of the testee rather than of the test itself.

Table 4.1 on page 62, shows both the way the operationalised components of CLA were intended to be sampled by the tasks, and the way the domain of CLA (i.e. topics, functions and conditions) were covered by the tasks. It is clear from this table that no component was neglected in the task design. Moreover, regular meetings of the EVA consultation team (referred to at the beginning of 'Speaking test specifications' pages 59 – 65) were held to discuss the task development, which provided a safeguard against poor sampling. However, in order to gain agreement from external sources that the tasks were perceived as sampling CLA adequately, a survey was carried out among teachers prior to distribution of the tests. The 30 teachers surveyed were attending one-day courses preparing them for using the EVA speaking material. The course began with a short module on 'what makes up speaking ability', in terms of components of ability and domain of language use. The teachers were then shown the actual test tasks, handed a blank form corresponding to that shown in Table 4.2 on page 68, and asked to put in crosses showing the way they perceived conditions, functions and components of CLA to be distributed across tasks or highlighted by them. The results of the survey are shown in Table 4.2, where the figures indicate the total numbers of crosses placed by the 30 teachers. The findings of this survey are fully discussed in this section, but are also relevant to the substantive aspects of validity discussed in 'The substantive aspect of validity' pages 71 – 74.

When compared with Table 4.1, which shows how the tasks were intended to give coverage to these topics, conditions, functions and components of CLA, the results shown in Table 4.2, showing how tasks were perceived to work, must be regarded as encouraging. As far as the conditions of speech are concerned, the teachers have overwhelmingly perceived the tasks as sampling the conditions in the way intended. In the case of functions, the teachers, not surprisingly, tended to perceive the tasks as highlighting the primary functions, such as imparting information or getting things done, rather than the secondary functions being carried out, i.e. being social, structuring discourse and check-and-repair. However, an analysis of the relative figures for each task shows that, on the whole, even the secondary functions have been recognised as being elicited by the tasks as intended. Task 1 is recognised as being associated with structuring discourse and task 2 with communicative check-and-repair. Little recognition was given to the function of being social (or 'socialising' as it was called in the original teachers' form). However, this may be due to the teachers' not interpreting the function in the way intended. This is not regarded as a problem, because the teachers recognised that pragmatic ability (the component most closely associated with being social)

Table 4.2: Coverage of topics, conditions, macrofunctions and components of CLA by tasks as perceived by 30 teachers

		task 1	task 2	task 3
topic	free selection			
conditions:				
purpose	describe	30	18	12
	narrate	17	4	10
	discuss	14	1	0
	instruct	0	29	2
	learn from instructions	6	29	5
	free selection	0	0	26
interlocutor	familiar	17	18	5
	unfamiliar	0	4	25
	peer	21	15	6
	adult	1	2	19
setting	school/test	26	18	5
	free selection	0	7	27
channel	face-to-face	29	27	23
	telephone	0	4	17
macrofunctions highlighted	imparting and seeking information	14	25	24
	expressing and finding out attitudes	17	3	3
	getting things done	1	20	22
	being social	3	3	5
	structuring discourse	10	6	9
	communication check-and-repair	6	14	8
component of CLA highlighted	microlinguistic	10	13	10
	textual	9	3	3
	pragmatic	8	14	20
	strategic	4	12	10

was activated by all tasks, but particularly task 3 (the role-play), as intended in the task design, and, to a lesser degree, task 2, where sensitivity is required in giving instructions.

The figures were generally lower for identifying the components of CLA highlighted than for the other parts of the form, and in fact several teachers made no attempt to fill this part in. This may be due to the fact that the concept being considered in this part – components of CLA – was the most unfamiliar and abstract of all. However, even here, the relative way the individual tasks have been recognised as highlighting components of ability is in line with what was intended in the design.

Thus the conclusion can be drawn that the tasks have truly been designed to sample components, and the evidence suggests that teachers perceive this to

be the case. However, as this evidence was not conclusive, particularly regarding the actual components of ability, it is worth carrying out future, a posteriori investigations, in a future round of testing, into how teachers perceive the way tasks sample language ability. Moreover, test transcripts could be used in future investigations to look for evidence of language ability in the performances on the different tasks.

UNCLEAR INSTRUCTIONS OR UNFAMILIARITY OF FORMAT

Since a condition of the test development has been that the methods should be innovative, the formats of both parts of the testing procedure are, at the outset, unfamiliar. Students are not used to being orally tested at all at this stage in their career. Teachers who have carried out oral testing are unaccustomed to testing in pairs using activity booklets, being more used to one-to-one interviews, often text-based.

Several measures have been taken to combat this unfamiliarity in the task format. A practice test has been issued, to be used as a classroom activity before the actual testing. Instructions on the test procedure are written briefly in students' booklets (Appendix A pages 268 – 273), and a script is provided for the director to tell the students exactly what to do (Appendix E pages 282 – 285). The teachers' handbook contains further advice on the whole testing procedure (Appendices D and F pages 280 – 281 and page 286). All instruction material has been inspected by teachers and by the EVA team of consultants, and trialled in schools with informal feedback. Moreover, centres have been contacted in all districts of the country, offering one-day courses to teachers embarking on the testing.

These measures have, one hopes, resulted in dispelling any lack of clarity in the instructions or unfamiliarity in the format. However, no systematic investigation has taken place among users as to how successful these attempts have been. It is therefore recommended that any future round of trialling be accompanied by questionnaires enabling both teachers and students to give systematic feedback on how easily they were able to follow the instructions and on how familiar the test was at the time of being carried out. This would provide the basis for an a posteriori investigation, possibly resulting in amendments to the instructions and training procedures.

TEST METHODS AND PROCEDURES THAT MAY PREVENT TESTEES FROM PERFORMING IN THE WAY INTENDED

Even if instructions are clear, methods and procedures can be such that they actually prevent the testees from doing the tasks intended. Great pains were taken, over numerous hours of first-hand trialling in schools, to arrive at methods and procedures that seemed to work well. Teachers gave valuable feedback through informal discussions, either individually or in the teacher group meetings that were held from time to time. This resulted in detailed

written instructions that cover, for instance, seating arrangements, timing, location and warm-up talk.

However, certain fundamental problems have arisen associated with the test methods used. Some of these have led to methods being abandoned, while others have been tackled so that the method could be retained. The two major stumbling blocks at the early stage were the use of role-play and the practice of testing in pairs.

When 'normal' role-play was tried (i.e. with both partners deciding what to say), it was found that it was impossible to keep the conversation on the intended course. This was overcome by introducing semi-role-play, wherein one partner's role is written down and the other is given a series of functions to perform, within a clearly defined context. This has largely been found to solve the problem, but has necessitated the use of two role-plays in task 3, in order that both partners have an equal opportunity to perform.

The aspect of the test method that was regarded at the outset as the most problematic was the fact that students were tested in pairs. It was pointed out that they might not get on, or that one partner might either dominate, or be so weak as to prevent any real communication. To combat the first of these hazards, teachers are given instructions that students must be paired on the sole criterion of being socially compatible. No restrictions are imposed regarding level of ability, as it seems undesirable to judge students before testing them. In order to ensure that no student can dominate, very rigorous roles are built into all the tasks, with both the director and the students themselves being given explicit instructions about who does what at any time (Appendices A, F and G pages 268 – 273 and 286 – 289). So that a very weak student cannot cause a breakdown, directors are told to step in to assist the student to give the essential information necessary to keep the task going, and to encourage the use of non-verbal communication, and even, as a last resort, Norwegian, so as to achieve this.

Whether or not this policy has worked can be investigated only a posteriori, on the basis of comparing grades given before testing and as a result of the test, for students paired with stronger and weaker partners. This investigation cannot, however, be carried out using the existing dataset; teachers tended to pair students with partners of a roughly similar ability, and there are only five cases where there was a difference of at least two grades between teachers' estimates of the students in a pair. A future trial should be carried out with a larger group of unequally paired students, analysing both grades and transcripts to test for partner effect.

TEST BIAS ASSOCIATED WITH CULTURAL BACKGROUND, BACKGROUND KNOWLEDGE, NATIVE LANGUAGE, ETHNICITY, AGE, OR GENDER

Norway is a relatively homogeneous society. However, a gradual influx of immigrants, particularly into the cities, has started to introduce a multicultural

element into Norwegian society. Moreover, the Sami people in the North of Norway have a very distinct culture and language. And the geography of Norway makes for very different lifestyles in its districts. With these factors in mind, as well as the gender issue, of which Norway is very conscious, the test was scrutinised at every point in its development for bias against any group.

The only easily identifiable grouping among the students is that based on gender. The next chapter reports on tests carried out on scoring data to see the extent to which bias in this area may be found to exist.

The SUBSTANTIVE aspect of validity

It was maintained in 'A validation framework' pages 30 – 31, that the substantive aspect of validity may be threatened by:

- tasks that do not fully engage the testees in the processes associated with the underlying theoretical model and domain of CLA
- tasks that essentially draw on processes that are irrelevant to the underlying theoretical model of CLA
- tasks that do not enable the testees actively to engage their language ability in a reasonably authentic way
- tasks that are uninspiring or off-putting and so fail to engage the testee in real communication.

TASKS THAT DO NOT FULLY ENGAGE THE TESTEES IN THE PROCESSES ASSOCIATED WITH THE UNDERLYING THEORETICAL MODEL AND DOMAIN OF CLA

In 'The content aspect of validity' pages 66 – 71, it was argued that the test tasks appeared (by design and with some teacher corroboration) to elicit a true sample of the language associated with the operationalised model of CLA, as outlined in 'Summary' pages 55 – 57. The same argumentation underlies the claim here that the processes associated with this operationalised model of CLA are, at least by design, sampled fully. This means that testees are required to 'do' all the different types of things described in the operationalised model. Thus a competent testee will spontaneously access and use fairly accurately the vocabulary associated with topics selected randomly from the domain of CLA. S/he will be able to adapt to the situational context and make the speech cohesive within and across turns. The speech acts associated with any of the six macrofunctions will be performed in a way that is fairly typical of the way native speakers perform these speech acts.

On the surface this may look identical to what was outlined on pages 66 – 71. However, the difference lies in the doing as opposed to the saying. It would be possible for testees to produce language indicating certain aspects

of CLA, without actually going through the processes of putting language ability to use as implied in 'Summary' pages 55 – 57. For instance, if testees have time to prepare for a test in advance, the words and expressions they use will not be produced spontaneously. And if the tasks entail very contrived situations, such as a testee telling a tester what s/he manifestly already knows, then the functions intended in the task may not actually be being performed. In the current test, no rehearsing or preparation time is allowed (although in the semi-role-play, some thinking time is built in), so that spontaneity is a feature of the test. The extent to which tasks are contrived or authentic will be returned to below.

TASKS THAT ESSENTIALLY DRAW ON PROCESSES THAT ARE IRRELEVANT TO THE UNDERLYING THEORETICAL MODEL OF CLA

In several documents, Messick cites two major sources as underlying test invalidity: construct-under-representation and construct-irrelevant variance (1989, 1994, 1995, 1996). Construct-under-representation occurs when the ability being tested is not fully sampled, i.e. the test is too narrow. Messick (1994) explicitly links authenticity with guarding against construct-under-representation in testing, in line with the stand taken here. Construct-irrelevant variance occurs when characteristics of the test takers other than the ability being measured consistently interfere with test scores, i.e. the test is too broad. This may typically occur in a test of listening that involves a high degree of reading ability. It can also occur because of features in the format such as multiple-choice items, which can induce an element of 'testwiseness'. Messick (1994) explicitly calls for directness as a means of guarding against construct-irrelevant variance. Messick (1989) also highlights the need for multiple tasks in the measurement of any construct, to guard against both major threats.

Since the current test is direct (i.e. the tasks simulate the way ability is normally used) and is composed of several tasks, each of which is 'scored' on the profile form, I feel that the test, in its design, should minimise the danger of construct-irrelevant variance. In other words, no processes other than those normally engaged in when taking part in spoken interaction, should consistently affect the performance. (Questions such as how far smiling or using body language to show interest should affect the performance are not fully taken up here. While these processes are not directly associated with CLA as it has been defined here, they cannot be said to be entirely irrelevant, they undoubtedly belong to communication, and may even be considered a non-linguistic extension of pragmatic ability.)

TASKS THAT DO NOT ENABLE THE TESTEES ACTIVELY TO ENGAGE THEIR LANGUAGE ABILITY IN A REASONABLY AUTHENTIC WAY

A test is always a test, and communication in test tasks will always be stamped by this fact. However, there are ways of designing tasks so that the communication is more likely to resemble that of everyday life. These include reducing stress, e.g. through making sure the input is readily comprehensible (Bachman 1990: 318) and through giving the testees a real purpose to communicate (Brown and Yule 1983b: 111). If the testees are fairly relaxed and have a genuine need to communicate, then the process of communication will take place in a reasonably authentic way. The fact that, at times, students are having to 'make believe' does not necessarily act as a barrier to normal communication processes. Cook (1997) points out that many natural conversations between adults are largely 'play and banter', and 'are often fantasies – not about the real world, but about a fictional one, in which there are no practical outcomes' (1997:230).

In order to enable test takers to communicate fairly naturally, two features have been incorporated into the EVA speaking test. The first is the fact that students are tested in pairs. In this way it has been possible to create a 'game' element, to the extent that both talking and listening are necessary in order to manage the tasks successfully. This is largely the case with tasks 1 and 2. The second feature is the use of pictures. By using comic strips for the narrative in task 1 and visual prompts for the instructions in task 2, the burden of understanding is greatly eased. In task 3 the student who improvises the conversation is blatantly acting a role, but entering this role is facilitated by the fact that s/he is always expected to be her/himself (i.e. never the adult) in a plausible situation for Norwegian teenagers when called upon to speak English.

Thus, the tasks have been designed so that students, can as far as possible, go through the processes of, for instance, telling, asking, clearing up misunderstandings, expressing views and even saving face, in an authentic way, because they genuinely need to and are able to relax sufficiently to do so.

Without getting inside the minds of the test takers, however, it is impossible to know the extent to which the tasks are succeeding in this respect. In the absence of any preliminary empirical investigation into how far students feel that they are communicating in an authentic way, it is suggested that a future a posteriori investigation should be carried out into this matter.

TASKS THAT ARE UNINSPIRING OR OFF-PUTTING AND SO FAIL TO ENGAGE THE TESTEE IN REAL COMMUNICATION

In the discussion above on authentic use of language, it was pointed out that the EVA tasks have elements of a game in order to involve both partners at all times. Additionally, an element of humour is introduced through the cartoon-

type pictures. Moreover, topics chosen for discussion are relevant to teenagers in Norway and the discussion is always personalised. Pains have therefore been taken to safeguard against uninspiring or off-putting tasks, and impromptu comments from students, recorded at the end of some of the tests, suggest that they find the tasks quite enjoyable. However, no systematic investigation has been carried out into students' reactions, and it is recommended that these reactions should be surveyed, a posteriori, in a future round of testing.

The STRUCTURAL aspect of validity

It was maintained in 'A validation framework' pages 30 – 31, that the structural aspect of validity may be threatened by:

- scoring procedures that do not fully and clearly reflect the specified model and domain of CLA
- clustering and division of constructs in the scoring system that are not supported by primary or secondary empirical evidence
- the creation of band-scale descriptors at different levels of ability that are not supported by some primary empirical evidence.

SCORING PROCEDURES THAT DO NOT FULLY AND CLEARLY REFLECT THE SPECIFIED MODEL AND DOMAIN OF CLA

In order to test whether scoring instruments fully reflect the model and domain of CLA, it is necessary to see whether all elements of behaviour defined as representing evidence of CLA in 'Operationalising components of CLA' pages 49 – 55, are included, explicitly or implicitly, among the elements addressed in these instruments, as described in 'Specifications for scoring procedures' pages 62 – 65. This is done component by component, and both the performance profile (PP) (Appendix B pages 274 – 276) and the band scales *message and fluency* (M/F) and *language structures and vocabulary* (L/V) (Appendix C pages 277 – 279) are examined. An analysis of these instruments is presented on pages 62 – 65.

While the performance profile is detailed by design, homing in on particular aspects of performance, the band scales are meant to be compact enough to be used for assigning level 'at a glance'. This means that different aspects of performance, such as politeness and the use of conventional expressions, may be referred separately to on the performance profile but summed up in a single reference, e.g. to appropriate style and politeness, on the band scale. Thus, while fairly explicit reference to aspects can be expected of the performance profile, only implicit coverage is normally expected of the band scales.

Microlinguistic ability is defined in 'Summary' pages 55 – 57, as implying being able to spontaneously:

- access and use with some accuracy the stock of words and expressions necessary to put the skills specific to speaking into practice – these being largely common smallwords and formulaic expressions
- access and use with some accuracy the stock of general and specialised vocabulary and language structures to talk about and 'operate within' the specified topics and associated situations
- perform functions of all the six specified macro-types, in a straightforward, transparent way
- produce sounds and intonation patterns well enough to allow the message to come across in full.

The last three of these elements can be regarded as covered, in principle, by the statements in both PP and band scales: structures and vocabulary are referred to quite explicitly (PP and L/V); performance of functions is implied by references to task achievement (PP and M/F); and the quality of phonological features is used as an adjuster alongside the band scales, as described in 'Specifications for scoring procedures' pages 62 – 65. However, the first element – the ability to use smallwords and formulaic expressions – is not mentioned or implied in either instrument.

Textual ability is defined on pages 55 – 57 as implying being able to spontaneously:

- produce turn-internal cohesion in such texts as descriptions and narratives, by ordering information conventionally, and by using links, such as pronouns and deictics (e.g. *over there*) and organising devices such as smallwords acting as discourse markers (e.g. *well* and *right*)
- produce across-turn cohesion in such texts as instructions and discussions, by the use of common conversational routines, and through smallwords acting as interactional signals (e.g. *okay* and *right)*
- speak smoothly, i.e. at a rate that is not detrimental to communication and without excessive hesitation. This smoothness may be assisted by formulaic expressions, verbal fillers (e.g. *you know*) and vague language (e.g. *sort of*).

The last of these elements is addressed in the M/F band scale, where flow and hesitance are mentioned, although no reference is made to the devices that may assist smoothness. Smoothness, or lack of excessive hesitation, is not given any mention in the PP, although the ability to 'keep going' is referred to. The two former elements, concerning cohesion, are not fully expressed. In the PP, some attention is given to within-turn structuring (i.e. the linking of ideas in specific tasks), but none to across-turn structuring, or to linguistic

devices that bring about coherence. The general statements in the M/F band scale concerning linking do not sufficiently convey the notion of two distinct types of cohesion – within and across turn.

Pragmatic ability is defined in 'Summary' pages 55 – 57, as implying being able to spontaneously:

- perform functions of all the six specified macro-types, in a conventional, 'idiomatic' way
- pay heed to the interlocutor's face, e.g. through the conventional use of empathisers, politeness expressions and hedges
- use transactional routines according to the purpose and setting of the speaking situation
- adapt language according to the age and familiarity of the interlocutor
- communicate by telephone as well as face to face.

Particularly through the series of questions relating to task 3, the performance profile gives quite full and explicit coverage to these elements in principle, although without specific reference to the devices used to take account of face. The L/V band scale's general references to idiomaticity, appropriateness of style and degree of politeness may be regarded as summarising these elements.

Strategic ability is defined on pages 55 – 57 as implying being able to spontaneously:

- use communication strategies that primarily employ English (i.e. analysis-based) only resorting to other (control-based) strategies as long as these do not involve using non-English forms
- carry out self-repair, check understanding on the part of the interlocutor, and indicate own lack of understanding, using the (small)words and other expressions normally employed to carry this out.

The explicit statements concerning non-use of Norwegian on both the PP and the L/V band scale may be regarded as summarising the first of these elements. However, the second, relating to check-and-repair, is not covered in either instrument.

To summarise, microlinguistic and pragmatic ability are fairly well covered in both instruments, while textual ability is very thinly covered by the band scales and the performance profile. Strategic ability is only partially covered. Common to most of the neglected areas is the lack of reference to linguistic devices necessary for achieving such essential qualities as smoothness of flow, coherence, face-saving and communication check-and-repair. These devices are context specific to some extent, but largely belong to the body of language already introduced as the smallwords of speaking (see Chapter 1 pages 1 – 6).

No attempt so far has been made to link individual components of CLA directly with either band scale, since it is not inherent in the test design that separate scores should be given on components, singly or in groups. This is consistent with the statement in the opening section of Chapter 3, pages 33 – 57, that components should not be regarded as independently measurable. However, a comparison between the nature of the elements making up the components of CLA as they are defined here and the nature of the two band scales cannot fail to suggest certain associations.

The *language structures and vocabulary* (L/V) scale contains references that are quite strongly associated with elements of both microlinguistic and pragmatic ability. The *message and fluency* (M/F) scale, as its name suggests, has to do with getting the message across in a reasonably fluent way.

No definition of fluency has yet been seriously offered, since this is to be taken up as the subject of Chapter 6 pages 122 – 156. However, in Section 1, Bygate's (1987) claim is presented that speaking involves, over and above an understanding of the system of the language, 'making decisions rapidly, implementing them smoothly, and adjusting our conversation as unexpected problems appear in our path' (1987: 3). For the time being, this extra 'thing' involved in speaking will be regarded as a working definition of fluency.

Getting the message across is largely dependent on what Bachman and Palmer (1996: 62) refer to as the personal characteristics that interact with language ability. However, it would be unfortunate if the *message and fluency* scale measured little more than personal characteristics. Fluency, as the term is used here, besides obviously depending on a certain basic microlinguistic ability, requires the ability to effect cohesiveness in communication, to play for time and to cope with potential breakdowns, and, as such, is clearly associated with the elements defined as making up textual and strategic ability. Since it is these two components of ability that have been shown to be the most seriously neglected in the band scales, it seems reasonable to assume, therefore, that the principal defects lie in the M/F rather than the L/V scale.

Besides being sufficiently comprehensive, band scales should, moreover, be clear and avoid vague concepts. North (1997: 441) underlines the need for 'definiteness' in band-scale descriptors. While such notions as vocabulary, grammar, pronunciation and even idiomaticity may be clear enough to be interpretable in the context of speaking, other notions, such as coherence, flow and hesitation, may require exemplifying. Moreover, as Fulcher (1993, 1996) demonstrates, various kinds of hesitation exist, which affect fluency in different ways. The subject of just how vaguely the EVA raters perceived the wording of scale descriptors and profile statements to be is taken up in 'Inter-rater reliability, pages 104 – 110.

Thus the scoring instruments have been exposed as being neither full nor clear, particularly in the case of the band scales, and most specifically the scale

relating to *message and fluency*. Any immediate recommendation of adjustments to the instruments is suspended, however, until further, a posteriori, investigations have been carried out on how well the instruments appear to have worked (on the basis of rater reliability) and on what transcripts reveal about the concept of fluency and the role of smallwords in bringing this about.

CLUSTERING AND DIVISION OF CONSTRUCTS IN THE SCORING SYSTEM THAT IS NOT SUPPORTED BY PRIMARY OR SECONDARY EMPIRICAL EVIDENCE

By the clustering and division of constructs we mean the way that statements referring to the various components of language ability occur jointly or separately for rating purposes. This clustering and division is discussed here with respect to the band scales only, where the statements are grouped, or clustered, into two scales, with a rating being given on each scale. All the statements on the performance profile, on the other hand, are independently rated. Statements in the band scales can be summarised as relating to the following elements:

MESSAGE AND FLUENCY

- initiative and contribution
- managing to put across the essentials of the task
- flow and hesitance
- ability to keep going without help
- linking and cohesion

LANGUAGE STRUCTURES AND VOCABULARY

- accuracy, variety and independence in language choices
- adequacy of vocabulary
- idiomaticity in language choices
- appropriacy in style and degree of politeness
- degree of coping without recourse to Norwegian.

The degree of comprehensibility of pronunciation and intonation as support to the message is assessed 'outside' the band scales, and this assessment is used as a final adjuster of the overall grade.

Put very crudely, while the *language structures and vocabulary* (L/V) scale is concerned with the language resources available to a person, the *message and fluency* (M/F) scale is to do with using these resources to initiate and maintain communication, getting the relevant message across as smoothly as possible.

As was explained in 'Specifications for scoring procedures' pages 62 – 65, this division and clustering of elements is based primarily on the conventional,

intuitive division fluency/language which is more or less observable in most oral assessment systems. Fulcher (1993) comments that in the many attempted component rating scales, two of the components are 'invariably the constructs "fluency" and "accuracy"'(1993: 122). This practice of separating fluency from some kind of knowledge and control of the language system was already established in the 1980s, e.g. in the American FSI (Foreign Service Institute) scales. It is evidenced in recent documents, such as the Council of Europe's *Common European Framework of Reference* (1996), where 'criteria for scaling aspects of communicative proficiency' (1996: 132) are shown in Table 4.3:

Table 4.3 Criteria for scaling aspects of communicative proficiency (Council of Europe 1996: 132)

PRAGMATIC	LINGUISTIC
(Language Use)	**(Language Resources)**
- Fluency	- General range
- Flexibility	- Vocabulary range
- Coherence	- Grammatical accuracy
- Thematic development	- Vocabulary control
- Precision	- Phonological control

Strategies (turn-taking, cooperation and asking for clarification) and socio-linguistic competence are regarded as separately scalable in the Council of Europe model. However, the fact that the elements grouped under 'language use' and 'language resources' in a document so influential in current European test development resemble fairly closely those in the EVA M/F and L/V scales is regarded as evidence for the structural validity of these scales. The use of two scales only in the EVA system is largely to do with the wish to help teachers without overburdening them with multiple scales, and with the fact that teachers already have an intuitive feel for the fluency/language divide, although, as will be shown in Chapter 6 pages 122 – 156, just what goes into the notion of 'fluency' is by no means universally agreed upon.

The inclusion of appropriacy, conventionality and idiom, representing pragmatic ability in the EVA L/V band scale, is based on the feeling that this ability is heavily dependent on what is referred to in the Council of Europe scaling as general and vocabulary ranges, i.e. language resources.

The element of 'not resorting to Norwegian' (associated with strategic ability) was placed on the *L/V scale*, originally because it seemed also, on the surface, to be largely dependent on linguistic resources, e.g. the necessary vocabulary to express a concept in English. However, in the previous

sub-section, it was argued that strategic ability, i.e. the ability to cope with potential problems in communication, is associated with fluency. In 'Operationalising strategic ability' pages 53 – 54, strategic ability was operationalised as including the use of communication strategies that do not involve non-English forms. Thus there is a clear place for the element 'not resorting to Norwegian' in the band scales. However, it follows logically from the way fluency has been discussed here that the place for this element is in the *message and fluency* scale, rather than the *language structures and vocabulary* scale.

The reason for grouping 'message' together with 'fluency' in the title of a band scale was based on a desire to emphasise that this dimension of ability should involve more than simply 'flow'. Fluency, as will be demonstrated in Chapter 6, tends to be associated, by raters, largely with smoothness, keeping going, and lack of hesitation. However, a student may keep going at length and say very little that is comprehensible, meaningful or relevant. The inclusion of message elements (taking the initiative, making a contribution and getting the essentials of the task across) is intended to ensure that the flow of speech – the ability to take the turn and keep going fairly smoothly – is judged in the context of what is required by, and relevant to, the task in hand.

The decision to place pronunciation and intonation outside the band scales was arrived at on the advice, through personal communication, of De Jong (1998). In his research into the relative contribution of constituent sub-skill scores at different levels of oral proficiency, De Jong (1991) examines the scores across a number of aspects of oral proficiency in French (after four to five years of study) of 25 subjects at upper secondary school in the Netherlands. Pronunciation (judged on 'read aloud' performance) marks itself out in De Jong's data as correlating very weakly with other measures of oral proficiency. De Jong maintains that this 'seems to be caused mainly by lack of development on this aspect for the upper two thirds of the distribution of subjects on the global ratings' (1991: 32). As it can reasonably be assumed that the English oral proficiency of the EVA test subjects (also having been exposed to four to five years of study) is generally as good as the French oral proficiency of De Jong's subjects, this lack of development of pronunciation can be taken to be as pertinent in the case of the EVA testees.

De Jong concludes that his data suggests that 'attainments along three dimensions of oral proficiency should be reported separately: pronunciation, accuracy and fluency. These dimensions are positively correlated, but an overall score, be it analytic or holistic, constitutes an unwarranted simplification' (1991: 34).

It seems reasonable to conclude, therefore, that a certain broad base of support can be found in conventional and current testing practice for the division and clustering within (and outside) the EVA assessment band scales.

The proviso is added, however, that any elements associated with strategic ability, such as the avoidance of using Norwegian, should be placed in the *message and fluency scale*.

Whether or not this clustering of elements in the band scales is further supported by primary evidence is the subject of 'The structural aspect' pages 113 – 117, where the factor analysis of scoring data is reported on.

THE CREATION OF BAND-SCALE DESCRIPTORS AT DIFFERENT LEVELS OF ABILITY THAT ARE NOT SUPPORTED BY SOME PRIMARY EMPIRICAL EVIDENCE

In order to discuss the extent to which the band-scale descriptors are supported by empirical evidence, it is necessary to describe the process by which they were built. This process was greatly facilitated by the fact that the school grading system was used as a basis for the EVA grading system. The decision to link these two systems was made jointly with the Ministry of Education, and will not be taken up fully here; it is sufficient to say that it was rationalised by the fact that teachers' routine assessment was expected to be enhanced in using the tests by having descriptors for performance at the different grades they were awarding.

At the time of test development (1994), these school grades were on a five-point scale, descended from an earlier norm-referenced system. They had, in principle, become criterion-referenced, but in fact lacked any tangible criteria. In order to link these school grades to the EVA six-level system, the broad middle band was divided into two. In this way the scale could be regarded as roughly equal in interval, and, in fact, teachers were advised to think of the bands as equal in range of ability. This is more an issue in the context of statistical data testing than in the actual use of the scale. As Alderson puts it, 'Ultimately, since we are not creating an equal interval scale, what will matter is whether assessors can use the scales and agree on their understanding of the descriptions that define the levels' (1991: 82).

The first step towards compiling descriptors was a survey parallel to that which had been used earlier in the compilation of criteria scales for the EVA writing test. A questionnaire was sent out to over 20 English teachers of the age group being tested, together with a copy of an early version of the test (which teachers were not required to actually use). A set of 26 'statements' about performance were given, some relating to specific tasks, others to general performance. For each statement the teacher was asked to indicate, by ticking a box, at which minimum level of oral ability (using an adaptation of the school grading system) they intuitively believed the statement would be true.

The result was both disappointing and illuminating. Whilst in the case of writing tasks, considerable consensus had been reached, in this case there was none. Most teachers replied that they were unable to fill in the form, and those

who did differed widely in their answers. The conclusion was that intuitive teacher estimates of performance could not be used here, which confirmed the impression that teachers are considerably less confident about and capable of evaluating their students' oral ability than their written.

This led to the birth of the performance profile. An embryo performance profile was written, made up of the statements used in the teachers' survey. The (then five-task) test was trialled in a local school, with ten students whose abilities spanned the full range of grades (except the absolutely lowest grade, which is virtually never given in speaking). Recordings from the trialling were rated by five EVA project-team raters, using the embryo performance profile, who also gave an intuitive global score, using the adapted scale of one (low) to six (high) ('one' being virtually never given). The class teacher also contributed a global score, converted to the new scale.

On the basis of these joint evaluations, each student was given an average global score and rerated on the performance profile, using the 'average placing' on each profile statement. The placings for the students at different levels were then plotted for each statement, and a rough picture was thus obtained regarding how students' performance at different levels might be described in terms of each of the profile statements.

An initial basis for describing students' performance at the six levels of ability on the two scales was thus obtained. These scales were repeatedly refined through trials over a period of about one year, by discussing how they actually described what the raters were hearing in the speech of students at different levels, and by adjusting them until they were agreed to be giving a fairly true picture. The use of two six-band scales made this way seemed in harmony with North's (1992) maxim 'have enough levels for learners to be able to see progress, to stimulate motivation, and for low level attainment to be credited [... [and]...] be user-friendly' (1992: 158).

Thus, the scale descriptors at the a priori stage can be regarded as being supported by some primary empirical evidence. However, further investigation needs to be carried out a posteriori, using the scoring data on a larger group of students than the initial ten, to see what kind of average profiles emerge for the different levels, and how this supports the existing scale descriptors. Furthermore, test transcripts should be used for comparisons of what is stated in descriptors with the actual performance of students at different levels. These investigations will not be carried out in the present study, but are recommended as subjects for future research, using revised, and I trust more reliable, scoring instruments to rerate the recorded and transcribed EVA speaking-test data. As a prelude to this research into descriptors, the investigation using transcripts in Part Two will provide some empirical evidence as to what might be written in the descriptors at different levels of a revised *fluency* band scale.

The GENERALISABILITY aspect of validity

It was maintained in 'A validation framework' pages 30 – 31, that the generalisability aspect of validity may be threatened by:

- methods and procedures for testing that are unclear or weakly defined, so that inconsistencies may occur in the way the test is carried out
- scales or other scoring instruments that are couched in vague terms (and so give rise to different rater interpretations)
- instructions and procedures for scoring that are unclear or weakly defined
- lack of rater training (so that scoring is potentially inconsistent)
- the influence of a weak or dominant partner.

Additionally, it was noted that invalidity with respect to the content, substantive, structural and external aspects would also undermine generalisability. These are dealt with in the relevant sections. In this section, only the listed threats, which concern reliability, will be examined.

METHODS AND PROCEDURES FOR TESTING THAT ARE UNCLEAR OR WEAKLY DEFINED, SO THAT INCONSISTENCIES MAY OCCUR IN THE WAY THE TEST IS CARRIED OUT

The methods and procedures were discussed in 'The substantive aspect of validity' pages 71 – 74. Here it was pointed out that detailed instructions on how to carry out the testing are given in the teachers' handbook. These instructions (presented in Appendix F on page 286) seem to support the claim that precautions were taken a priori to prevent inconsistencies from occurring in the testing procedure. The instructions laid down in the students' booklets and the test director's script (Appendix E pages 282 – 285) are intended to exercise strict control over the actual methods. However, a future check should be carried out a posteriori, using transcripts to examine the extent to which the methods and procedures are actually adhered to in practice.

SCALES OR OTHER SCORING INSTRUMENTS THAT ARE COUCHED IN VAGUE TERMS

It has already been pointed out, in 'The structural aspect of validity' pages 74 – 82, that the terms used in the scoring instruments are often abstract and lacking in actual linguistic exemplification. Just how great a problem this is can be gauged by seeing the extent to which raters were able to agree in their scoring on points of the performance profile. This will be the subject of a further, a posteriori validation using scoring data as well as raters' judgements of vagueness, covered in the next chapter.

INSTRUCTIONS AND PROCEDURES FOR SCORING THAT ARE UNCLEAR OR WEAKLY DEFINED

The above comments on the safeguarding of consistent testing procedure apply equally to the scoring procedures, which are written down in teachers'

material and shown in Appendix D pages 280 – 281. The inter-rater reliability testing of overall grades carried out a posteriori, and reported in Chapter 5 pages 96 – 120, should shed some light on the extent to which these procedures have, in fact, been safeguarded. Future examination of test transcripts should reveal more detailed information on the extent to which procedures have been adhered to.

LACK OF RATER TRAINING

In-service courses in the use of the whole EVA assessment system have been given to teachers and teacher trainers, who are supposed to pass on their expertise to colleagues. However, as it is unrealistic to imagine that this will reach all teachers, who will normally be the raters of their students, the teachers' material contains instructions for scoring (Appendix D) as well as a video showing the performances of six students, accompanied by filled-in profiles of their performances. The unfamiliarity of the format of the scoring instruments has been further addressed by the inclusion in the testing package of forms and checklists for assessment of performance during routine classroom activity (Appendix G pages 287 – 289). This material is designed to encourage continual documented assessment, including self-assessment, of speaking ability, besides familiarising teachers and students with the elements to be listened for when assessing this ability.

THE INFLUENCE OF A WEAK OR DOMINANT PARTNER

The possible influence of a weak or dominant partner was discussed in 'The content aspect of validity' pages 66 – 71, where it was made clear that measures have been taken to prevent it. The present dataset does not contain enough examples of pairs with a wide 'gap' in level to enable investigation at this stage. However, this important issue should be returned to in future research, focusing on the performance of students with partners of markedly different ability.

The EXTERNAL aspect of validity

It was maintained in 'A validation framework' pages 30 – 31, that the external aspect of validity may be threatened by:

- using external criteria that measure different abilities from the test in question
- failing to look for discriminant evidence to ensure that the test is not measuring unrelated abilities
- using external criteria whose validity is unknown.

USING EXTERNAL CRITERIA THAT MEASURE DIFFERENT ABILITIES FROM THE TEST IN QUESTION

The external criteria used for evaluating the test results were teachers' grades and students' self-assessment. As Norwegian teachers have not traditionally given grades for spoken language ability (although this is currently being introduced), an estimate of this ability was obtained as follows. Teachers were asked, at the outset of the piloting of all EVA test parts, to send the previous end-of-term grade of each student, i.e. a few months prior to testing, accompanied by an estimate of whether the grade would be higher or lower if awarded on the spoken (or written) language ability alone. The grade for speaking ability was accordingly computed either by using this grade as it stood or by marking it up or down one point (all scores being converted to the EVA six-point scale). This ensured, as far as it was deemed possible, that the scores recorded were measuring the ability being tested, i.e. speaking ability.

In the case of students, a self-assessment was obtained by the students' rating their likely performance on a scale of one to five on a series of tasks that closely resembled those given in the test (see Appendix H pages on page 290). This was done prior to testing. An average score was thus recorded for each student, which can be regarded as being based on speaking ability, as far as the general ability to cope with representative speaking tasks is concerned. The outcome of the external validation is reported in 'The external aspect' pages 100 – 103.

FAILING TO LOOK FOR DISCRIMINANT EVIDENCE TO ENSURE THAT THE TEST IS NOT MEASURING UNRELATED ABILITIES

In correlating the two sets of measurements, in the process described above, we are looking for evidence of convergence, or common ground being measured. However, it is also necessary to look for discriminant evidence to ensure that the test is not actually measuring something else, e.g. reading ability. Scores from the EVA tests of listening and reading, as well as teachers' estimates of writing ability are used to provide such discriminant evidence of validity, and reported on pages 100 – 103.

USING EXTERNAL CRITERIA WHOSE VALIDITY IS UNKNOWN

In the absence of the possibility of applying any other external criteria, the two chosen were the only ones available. However, each of these external criteria – teacher estimate and self-assessment – has been recognised as such, e.g. in the validation studies carried out on the ELTS (English Language Testing Service) speaking test, reported in Criper and Davies (1988), and on the IELE (Lancaster University's Institute of English Language Education) Placement Test, cited in Wall *et al.* (1994).

The CONSEQUENTIAL aspect of validity

It was maintained in 'A validation framework' pages 30 – 31, that the consequential aspect of validity may be threatened by:

- test tasks and methods that draw on irrelevant abilities
- scoring procedures that do not encourage the learner to assess his/her own performance
- lack of any analytic feedback on individual strengths and weaknesses
- band-scale descriptors that are vague or negative, so that they do not help learners to realise what they can and need to be able to do
- unclear instructions to users on how (and how not) to interpret test results
- failure to restrict inferences, made from test results, to what the testee can do, in the content domain specified.

Additionally, lack of validity with respect to the content, substantive, and structural aspects was found to threaten the consequential aspect (specifically by affecting 'washback'). These are dealt with in the relevant sections. In this section, only the listed threats will be examined.

TEST TASKS AND METHODS THAT DRAW ON IRRELEVANT ABILITIES

The EVA speaking test is direct. Moreover, it selects its samples freely from the topics, eliciting the full range of functions, under all specified conditions, as defined in 'The situation of the testees' pages 46 – 49. It does not draw on specific 'knowledge' of any topic. These factors should safeguard it to a great extent against rewarding aspects of performance that draw on irrelevant abilities, such as testwiseness or topic familiarity.

SCORING PROCEDURES THAT DO NOT ENCOURAGE THE LEARNER TO ASSESS HIS/HER OWN PERFORMANCE

Although students are given a self-assessment test on speaking (see Appendix H on page 290), rating their general ability to perform certain tasks, they are not directly involved in assessment of their test performance. It is recommended that students' booklets be amended to contain a self-assessment form similar to that issued for self-assessment on classroom tasks (Appendix H).

LACK OF ANY ANALYTIC FEEDBACK ON INDIVIDUAL STRENGTHS AND WEAKNESSES

Profiling of strengths and weaknesses is a strong feature of the EVA speaking test. This is at its most evident in the performance profiles, but the placing of performance on two band scales, with pronunciation and intonation also being taken into account, yields a further profile-like picture of performance. Moreover, the material issued for both teachers and students for assessing

classroom performance (Appendix G pages 287 – 289) not only contributes to the profiling of ability, but also encourages it to be sustained as an ongoing process.

BAND-SCALE DESCRIPTORS THAT ARE VAGUE OR NEGATIVE, SO THAT THEY DO NOT HELP LEARNERS TO REALISE WHAT THEY CAN AND NEED TO BE ABLE TO DO

The fact that the band-scale descriptors (Appendix C) are at times vague has already been suggested (see 'The structural aspect of validity' and 'The generalisability aspect of validity' pages 83 – 84) and will be further addressed as part of the a posteriori validation in the next chapter. However, the descriptors have been consciously designed to be positive as far as is possible, so that students can see, at each level, what they can do, and what they need to be able to do to reach the next level. Whether this is in fact clear to learners using the scales should be the subject of future a posteriori research, based on students' opinions.

UNCLEAR INSTRUCTIONS TO USERS ON HOW (AND HOW NOT) TO INTERPRET TEST RESULTS

It is made clear in the teachers' material that the main purpose of the tests is to give indications of strengths and weaknesses. It is pointed out that assessment should be ongoing in order that a true profile can be obtained. Grades given on the test should be interpreted with the help of the band scales, and it is emphasised that students themselves should be familiar with these scales. The following scale is provided for quick-and-easy interpretation of grades on all EVA tests:

grade 6 = performance is outstanding – very good on all factors
grade 5 = performance is consistently good
grade 4 = performance is, at times, good and beyond the basic
grade 3 = performance is adequate to communicate the basic message
grade 2 = performance is adequate to communicate some of the simplest message
grade 1 = performance is inadequate for any coherent communication

The extent to which users are interpreting test results as intended should be the subject of future research, based on responses from both teachers and students.

FAILURE TO RESTRICT INFERENCES, MADE FROM TEST RESULTS, TO WHAT THE TESTEE CAN DO IN THE CONTENT DOMAIN SPECIFIED

As the EVA speaking test is issued alongside tests in the other three macroskills of reading, listening and writing, it is unlikely that the speaking test result will be used to imply any ability other than speaking. And as the

content domain is defined, in principle, to include the full range of speaking tasks and contexts relevant for the students being tested, it can reasonably be concluded that inferences are being restricted to this domain.

Summary and conclusions

In this chapter, the EVA speaking test has been examined a priori, with respect to each of the six aspects in the framework for validation, worked out in Chapter 2 pages 9 – 32. Certain judgements have been made on the basis of the test as it stands, while some areas of validity remain to be investigated at a later stage in the study. Other areas are outside the scope of the present study but are recommended for future investigation. Moreover, the extent to which the test's tasks and scoring instruments reflect the model of CLA has been evaluated in the course of the validation process. Because it is fundamental to any judgement on the test's validity, this evaluation is summarised before the conclusions on validity are presented.

Conclusion on the extent to which the model of CLA is represented in the test

The conclusions reached on the extent to which the model of CLA is represented in the test in this chapter are summarised in Tables 4.4 a–d. Each individual aspect of performance making up one of the components of ability, as presented in 'Summary' pages 55 – 57, is marked for coverage in the tasks and in the two scoring instruments: the performance profile, and the band scales for *message and fluency* (M/F) and *language structures and vocabulary* (L/V). The tasks are presented in Appendix A pages 268 – 273 and the scoring instruments are presented in Appendices B and C pages 274 – 279, and analysed in 'Specifications for scoring procedures' pages 62 – 65. The judgement regarding tasks is based on whether the task, as it is designed, has the potential to elicit the particular aspect of performance concerned. This judgement is backed up to a large extent by the teachers' verdicts presented in Table 4.2 on page 68, although these dealt with components of CLA generally, and did not take account of the detailed aspects of performance making up these components. Judgements on the degree to which the scoring instruments reflect the model are based entirely on conclusions drawn in 'The structural aspect of validity' pages 74 – 82.

Tables 4.4 a–d illustrate the fact that the model of CLA is not inherently underrepresented in the language potentially elicited by the test tasks. However, gaps in the representation of the model through the scoring instruments are very evident. These gaps occur most notably in the case of textual and strategic ability, where the elements, as was pointed out in 'The

Table 4.4a The extent to which microlinguistic ability is represented in the test

	task	band scale	profile
access and use with some accuracy the stock of words and expressions needed to put the skills specific to speaking into practice – these largely being common smallwords and formulaic expressions	all	–	–
access and use with some accuracy the stock of general and specialised vocabulary and language structures to talk about and 'operate within' the specified topics and associated situations	all	L/V	yes
perform functions of all the six specified macro-types, in a straightforward, transparent way	all	M/F	yes
produce sounds and intonation patterns as well as other prosodic features well enough to allow the message to come across in full	all	adjuster	yes

Table 4.4b The extent to which textual ability is represented in the test

	task	band scale	profile
produce turn-internal cohesion in such texts as descriptions and narratives, by ordering information conventionally, and by using links, such as pronouns and deictics, e.g. *over there,* and organising devices such as smallwords acting as discourse markers, e.g. *well* and *right*	1	M/F partial	partial
produce across-turn cohesion in such texts as instructions and discussions, by the use of common conversational routines, and through smallwords acting as interactional signals, such as *okay* and *right*	2 and 3	–	–
speak smoothly, i.e. at a rate that is not detrimental to communication and without excessive hesitation. This smoothness may be assisted by formulaic language and verbal fillers, e.g. *you know* and vague language, e.g. *sort of*	all	M/F partial	–

structural aspect of validity' pages 74 – 82, would appear to have a natural home in any band scale associated with fluency. As can be seen in the tables, the current M/F band scale fully covers very few elements that are associated with components of CLA at all, while the L/V scale covers a considerable number. It can also be seen that elements neglected on the band scales tend also to be neglected on the performance profile. Thus the conclusion is reinforced that the model of CLA is reflected in the tasks but is only partially

Table 4.4c The extent to which pragmatic ability is represented in the test

	task	band scale	profile
perform functions of all the six specified macro-types in a conventional, 'idiomatic' way	all	L/V	yes
give regard to interlocutor's face, e.g. through the conventional use of empathisers, politeness expressions and hedges	all	L/V	yes
use transactional routines according to the purpose and setting of the speaking situation	mainly 3	L/V	yes
make appropriate language choices according to the age and familiarity of the interlocutor	mainly 3	L/V	yes
communicate by telephone as well as face-to-face	3	L/V	yes

Table 4.4d The extent to which strategic ability is represented in the test

	task	band scale	profile
use communication strategies that primarily employ English (i.e. analysis-based), only resorting to other (control-based) strategies as long as these do not involve using non-English forms	mainly 2	L/V *(move to M/F)*	yes
carry out self-repair, check understanding on the part of the interlocutor, and indicate their own lack of understanding, using the (small)words and other expressions normally employed to carry this out	mainly 1 and 2	–	–

reflected in the scoring instruments, which fail to give coverage to textual and strategic ability as well as to that part of microlinguistic ability that involves access to the smallwords of speaking. The source of this weakness can largely be traced to gaps in the *message and fluency* scale, as well as in the performance profile. The transference, recommended in 'The structural aspect of validity' pages 74 – 82, of any element associated with the use of non L1-based communication strategies to the M/F scale should go some way towards remedying this imbalance.

Conclusions on the validity of the test

The conclusions on how valid the test appears to be with respect to each aspect, and which imminent or future issues should be addressed, are summarised below. Before embarking on this, it must be emphasised that the

a priori validation process has been able to do no more than examine the potential validity of the test. The concern has been to find weaknesses that could undermine the test's validity. Whatever has 'passed the test' has simply done so by virtue of not actually constituting an inherent source of invalidity. However well designed the test appears to be, no pronouncement can be made that it is valid. Moreover, any conclusions on validity relate to the test as it stands today. Future versions of the test will differ in certain aspects, and these differences will have to be taken into account in drawing conclusions on the validity of such versions. Only use of the test, and scientific scrutiny of data emerging from this use, over years if necessary, can actually confirm or refute its validity.

THE CONTENT ASPECT OF VALIDITY

No obvious weaknesses have been found in the content aspect of validity of the test. In other words, the tasks appear to have been designed with the potential to elicit language products (i.e. things people say) that contain evidence of ability. The scoring data investigation will provide indications of the success or otherwise of the attempts to prevent test bias with respect to gender, which could endanger the generalisability of test results.

It is recommended that future investigations be carried out into teachers' opinions as to how well the tasks seem to elicit evidence of CLA. Moreover, feedback should be systematically obtained, in any future round of trialling, from both students and teachers, regarding the clarity of instructions and familiarity of the test format, as both of these facets can affect the extent to which the tasks are actually carried out as intended. It is also suggested that transcripts are examined in order to see how the methods and procedures actually used in testing reflect those intended to be used. Measures have been taken to counter any effect of a weak or dominant partner, but the extent to which these seem to have been successful is recommended as the subject of future research, using scoring data and transcripts for a wider group of students.

THE SUBSTANTIVE ASPECT OF VALIDITY

It appears that the design of the test should ensure the substantive aspect of validity, so that testees are actually able to go through processes that are representative of language ability in use, and do not significantly go through irrelevant processes. However, as it is impossible to know exactly what is going on inside the mind of a testee, a further survey is recommended into how authentically students feel they are communicating when carrying out different parts of the tasks, as well as how engaged they are by the tasks.

THE STRUCTURAL ASPECT OF VALIDITY

The structural aspect of validity, which concerns the extent to which the structure of the scoring procedure actually reflects the structure of language ability, was found to have apparent strengths and weaknesses.

The way constructs are clustered in the band scales has been found to draw support from conventional and current testing practice, insofar as the separation in the band scales of *message and fluency* and *language structures/vocabulary,* and the way elements are clustered within these scales, is concerned. Moreover, support is derived from secondary literature for the placing of pronunciation and intonation outside the scales. However, it is recommended that reference to the use of non L1-based communication strategies should be transferred from the L/V to the M/F scale. Factor analysis on profile performance scores should yield some primary evidence to indicate how far the fluency-language clustering and division of elements and the separate assessment of pronunciation and intonation are justified.

The band-scale descriptors at different levels of ability have been found to be supported by a small amount of primary evidence, and it is recommended that future larger-scale investigations should be carried out into this, using both scoring data and transcripts.

An inherent weakness in the scoring instruments is that they – and particularly the band scale relating to fluency – do not fully or clearly reflect the components of CLA, especially in the case of textual and strategic ability. This is largely ascribable to a lack of explicit reference to linguistic devices for bringing about fluency, notably through the use of smallwords. However, no immediate recommendations for amendment are made, pending, firstly, the results of the investigations in the next chapter, to see how well the instruments work in terms of inter-rater agreement, and secondly, the findings in Part Two on what transcripts reveal about the concept of 'fluency', and the role of smallwords in bringing it about.

THE GENERALISABILITY ASPECT OF VALIDITY

As explained in 'Six central aspects of validity' pages 29 – 30, the study of generalisability is restricted here to considerations of reliability. The fact that very detailed video-accompanied instructions are issued for scoring procedures, and that training is offered widely, contribute positively to the reliability of test results.

However, the rather vague and abstract terms in which the elements on the scoring instruments are couched appear to be a source of potential weakness. The inter-rater reliability studies carried out in the next chapter should indicate how much of a problem this is. Rater judgements on this vagueness will also be studied. The possible effect of a weak or dominant partner on performance should be investigated in the future, using additional data.

THE EXTERNAL ASPECT OF VALIDITY

The external criteria used for validation have been found to be good insofar as they measure speaking ability and are of the type (teachers' and students' estimates of ability) used in other studies and regarded as valid yardsticks. Having made these points, it remains to be seen, from the analysis of scoring data in the next chapter, just how valid the test scores are when measured against these external criteria.

THE CONSEQUENTIAL ASPECT OF VALIDITY

While findings pertaining to content, substantive and structural validity all have an impact on the consequential aspect of validity, as pointed out in 'The consequential aspect of validity' pages 86 – 88, certain additional points can be made that specifically relate to this aspect. It is seen as a serious flaw in the scoring procedure that learners are not involved in the assessment of their test performance. This reduces the potential of the test in raising learners' awareness, and hence their short- and long-term ability to help themselves. A recommendation is made that self-assessment forms should be placed at the end of the students' test booklets, along similar lines to those already issued for self-assessment on performance of classroom tasks.

Moreover, any conclusions on the actual long- or short-term uses and consequences of the test can be made only on the basis of future surveys. It is suggested that a series of questions be issued to the various test users in order to chart the actual ways testing is used, and its consequences.

OVERALL VALIDITY

An overview of the findings from this chapter on validity in the test is presented in Table 4.5 on page 94, which lists what has been judged to be satisfactory and unsatisfactory in the various aspects of validity. Recommendations are summarised on both what should be investigated in the remainder of this study and which future courses of action should be taken.

As can be seen from the table, all aspects of validity have been judged to be satisfactory in some respects, and in the case of certain aspects (notably content, substantive and external), no inherent weaknesses have been revealed at this stage. However, virtually all aspects are judged to be in need of future research. In the case of the structural, generalisability and consequential aspects, weaknesses have been revealed, and these are principally located in the scoring instruments – chiefly in the scale relating to message and fluency – and are largely caused by an underrepresentation of the model of CLA and an apparent lack of precision and of concrete, 'linguistic', references in the terminology of these instruments. These weaknesses have far-reaching consequences for the test's perceived validity. Not only do they immediately undermine the structural validity of the scoring instruments, but they could also potentially cause a negative washback effect, thus affecting the test's

Table 4.5 Overview of findings on the validity of the EVA speaking test

aspect of validity	judged to be satisfactory	judged to be unsatisfactory	to be investigated in this study	recommended for future study
CONTENT	tasks designed to elicit language products that contain representative evidence of ability		scoring data findings on gender effect	survey among teachers/students for clarity/familiarity survey among teachers on tasks' potential to elicit evidence of CLA examine transcripts and data to check procedures
SUBSTANTIVE	tasks designed to put testees through processes representative of language use			survey among students for response to tasks and perceived authenticity of communication
STRUCTURAL	clustering and division of constructs supported by secondary evidence band-scale descriptors supported by some primary evidence	components of CLA not fully and clearly reflected in scoring, esp. message/ fluency scale non-use of L1 to be moved to M/F scale	transcript findings on fluency and role of smallwords factor anal. profile scores to support the fluency/ language divide	search for primary evidence to support band-scale descriptors of performance at different levels
GENERALISABILITY (subsumes content, substantive, structural and external aspects)	rater training provided for operationalised test	apparent lack of precision and concreteness in the wording of scoring instruments	scoring data findings on inter-rater reliability raters' judgements on vagueness in scoring instruments	investigation of scoring/transcript data from a larger group to shed light on weak/dominant partner effect examine transcripts and data to check consistency of procedures
EXTERNAL	recognised external criteria used		scoring data findings on test scores/external measures	
CONSEQUENTIAL (subsumes content, substantive and structural aspects)		lack of self-assessment		surveys among test users on use and consequences of testing and appropriacy to new curriculum

consequential validity. The knock-on effects of inexplicitness in these instruments can also detract from rater reliability and hence from the generalisability of the test results.

Moreover, certain questions have been raised, in the course of this part of the validation process, that are as yet unanswered, but which the scoring data may elucidate. The content aspect of validity may be jeopardised by test bias with respect to gender, which must also be looked into. Moreover, in order to be able to claim that the test is 'working' as a test of speaking, i.e. yielding results that seem sensible, it is necessary to establish the external validity of the test.

Besides summing up the test's profile of validity, Table 4.5 on page 94, also lists areas of validity to be investigated in this and future studies. The important question pertaining to the structural aspect of validity, as to how fluency may best be described in the scoring instruments, can only be answered with reference to test transcripts and will be looked into in Part Two. Otherwise, the weaknesses and the unanswered questions outlined here are addressed in the next chapter, using the data – principally raters' scores – obtained during the first wide-scale round of piloting.

5 Validation based on scoring data

In this chapter, the validation process moves into a new phase, based on the data collected during the national trialling of the EVA speaking test during the spring of 1995. Statistical testing on this data is used to help shed light on aspects of the test's validity which were shown to be in need of further investigation and within the scope of the present study, as shown in Table 4.5 on page 94. The external, structural and generalisability aspects of validity are examined. The chapter concludes by drawing together the findings to reappraise the test's validity, and to identify any need for improvement.

Data and methods

Before proceeding to give an account of the dataset and methods of analysis it is necessary to place the data collection in context. This context was primarily the trialling of a set of tests (of which speaking was only one) in a short space of time, by a number of schools across Norway, at the request of the Ministry of Education. Teachers were under pressure, not only of time and the physical and timetabling challenges of testing, but also of justifying the extra activity involved to parents and others concerned. It was necessary to ensure the maximum benefit to those immediately involved, with the minimum disruption.

Teachers were encouraged to volunteer to take part in trialling the speaking test, in the knowledge that they would get feedback on the performance of a cross section of their students, and would, moreover, be introduced to a new form of oral testing. To minimise the pressure on teachers, and because a primary concern was to see how the test material 'worked' in the hands of people familiar with its design, the actual testing was largely done by myself, with class teachers only doing it in some more far flung regions of the country, following detailed instructions. Members of the EVA Project Group were drafted in as raters.

An aim was to develop routines that any class teacher could follow at his/her own convenience, from administering the test to giving informal grades and feedback on strengths and weaknesses, with the help of a video and instructions, and where possible backed up with local training sessions. This was not to be a high stakes test; it would not be feasible to expect that two

raters would be used as a norm, although this was encouraged; nor was there any question that statistical analysis would be routinely carried out on scores.

However, as validation of the test was perceived from the start as an integral part of the overall trialling, every effort was made to collect data in a way that would allow this, while not putting an impossible strain on those involved. Care was taken to collect high-quality recordings from a random set of as many students as possible, given the logistic and budgetary restraints, with an even geographical distribution across the country. These were all rated by the tester and one other person, with the final grade being the average of the two ratings. Although rater training as such was not given, those brought in as second raters were chosen for their prior involvement in the design stage of the test and their experience as English teachers/teacher trainers at this level. Not only were these measures to be carried out in the interests of the research, but they were also intended to safeguard the fairness of the assessment of students' performance.

In retrospect, it is clear that other precautions should have been taken. A trial run with all raters using a number of tapes should have been undertaken; the tapes should have been distributed to ensure similar spreads of level as far as possible – this would prevent 'truncated' samples being graded by individual raters. Ideally, too, multiple rating should have been carried out with a number of tests, making it possible to compare leniency/severity, and through multi-faceted Rasch analysis, to identify more precisely the effects of different raters than has been the case.

Having made this point, however, it must be argued that the extra burden imposed by multiple rating on the individual raters in terms of the quantity of tapes assessed (each test taking around 30–40 minutes to rate) would have been unacceptably high in proportion to the actual value of the additional data obtained. McNamara (1996: 127) cites two purposes for modelling rater characteristics by multi-faceted measurement. One is to enable the grades of candidates to be adjusted to compensate for rater effect. The other is to provide a basis for research into rater differences. The former reason clearly did not apply in the case of the kind of test being developed here. And while rater differences were highly relevant to the issues of generalisability being studied here, no in-depth study of rater behaviour of the type cited by McNamara was considered feasible, given the small pool of raters and the one-shot nature of the rating.

It is therefore hoped that the reader will be forbearing during the following account; the statistical tests used are unsophisticated, being mainly of a simple, inter-rater correlation type. This is owing to the size and nature of the dataset, which was decided largely by considerations of expedience and appropriacy both to the immediate test trialling situation and to the type of testing envisaged in the fully-operational test. This having been said, however, the bottom line must be that statistics have been used here to give indications

of what seems to be the case, and to support what common sense, convention and theory have already suggested.

The data used in this investigation is made up of various ratings of the speaking ability of students who took the EVA speaking test in the national trialling. Students were tested in pairs in ten schools, which were selected randomly for oral testing from within the 47 schools in the wider EVA testing project. Six to eight students from each school were picked alphabetically and then paired by the teacher according to the sole criterion that they 'got on' together (see 'The generalisability aspect of validity' pages 83 – 84, regarding the possible implications of this). The recorded performances were rated by the tester (normally myself (rater 1), but occasionally the class teacher), and at least one of the other four EVA raters (raters 2 to 5). The rating was carried out in stages. Firstly, each rater listened to the recorded test and filled in a performance profile for each student (see Appendix B pages 274 – 276). On the basis of their observations, the raters were then asked to 'place' the performance on two band scales (see Appendix C pages 277 – 279), then to award a single overall grade on the basis of these scale placings, following strict guidelines. This procedure is described in greater detail in 'Specifications for scoring procedures' pages 62 – 65.

Two datasets were actually used in the analyses: one was based on a sample of 59 students, and contained teachers' estimates, overall test scores, performance profile scores and gender information. This dataset used the *Statview* statistics programme. The other, which used *NSD Stat* programme, contained a larger dataset, drawing on 554 students, 70 of whom took the EVA speaking test. However, this dataset was more limited in the range of relevant information it contained, this being scores on all parts of the EVA test (i.e. reading, listening, writing and speaking), teachers' estimates and gender information. The following data were used in the analysis:

- scores on a 22-point performance profile, by at least two raters
- overall evaluation of performance on a six-band scale, by at least two raters
- teachers' (pre-test) estimates of students' oral skills
- students' self-scores in speaking, calculated from results from the EVA self-assessment questionnaire (see Appendix H on page 290)
- students' gender information.

The dataset was made up almost completely of scores/grades and other information that was returned by those involved in the piloting of the test. This has been advantageous in that the data was 'genuine', being obtained under the conditions of actual use of the material. However, as has been pointed out above, it has been a weakness insofar as the data was restricted to what was available at the time. Although contact was maintained with the raters, no large-scale supplementing of test-score data was possible. It was possible,

however, to ask raters certain general questions, notably how vaguely they believed statements in the scoring instruments to be worded. Regrettably though, it did not allow for the provision of one very valuable set of data, i.e. the individual ratings on the two separate band scales. At the time of testing, raters were required to return the performance profile form and the global grade only, as these were regarded as the two main 'outcomes' of the testing. This lack of information must be regarded as a weakness inherent to the dataset. The way raters have assessed aspects of performance covered by the band scales has had to be judged on the basis of the corresponding sub-skill scores on the profile form (grouped according to whether they are language- or fluency-related).

In the statistical analyses, only parametric tests have been used. Such tests use absolute values of data and yield more information than non-parametric tests, but require certain conditions to be fulfilled (a normal distribution and 'equal interval' data). Neither EVA grades nor scores truly qualify as equal interval scales (i.e. the difference between one score/grade and the next is not guaranteed as constant) and, therefore, do not strictly merit the use of parametric statistical analysis.

However, parametric testing has been used in order to widen the scope of the information yielded. This is felt to be justified by the fact that, firstly, in the case of grades and teachers' estimates, a near-normal distribution existed , and secondly, raters were asked to regard scales as equal interval, the scales having been designed to approximate to the equal interval ideal. Data values have thus been entered as integers, and treated as if they were equal interval, being adjusted with decimal figures to account for teachers' qualifying comments and 'in-between' grades.

The a posteriori validation process

As Table 4.5 on page 94 shows, three aspects emerged from the validation procedure in Chapter 4 pages 58 – 95, as being in need of further investigation within the present study: structural, generalisability and external. The investigations cited in the table have, with one exception, involved analysing the data listed above, and are reported on in this chapter, with the aspects being considered in the reverse order to that cited above. The remaining investigation, involving the nature of fluency and the role of smallwords, has been carried out using transcripts and will not be dealt with until Part Two.

The a posteriori validation based, covered in this chapter, begins by examining the degree of external validity that the test appears to have, giving a preliminary indication of whether or not the test is working as it should. If the test is not apparently working, there may be little point in proceeding with the investigation! Next, test bias is looked for, potentially affecting the

content aspect. Generalisability is then examined, considering the question of inter-rater reliability. The results of this part of the investigation give an idea of how much credence can be given to individual ratings. The structural aspect of validity is then opened up, with the help of factor analysis on raters' scores on individual sub-skills, to see whether the way in which these scores cluster lends support to the way elements are clustered in the test's band scales. In the conclusion to the investigation, the consequential aspect of validity of the test is briefly considered, with recommendations being made on the basis of this a posteriori validation, with washback in mind.

The EXTERNAL aspect

As indicated in Table 4.5 on page 94, the external aspect of validity is investigated here with respect to the comparability between test scores and external measures of speaking ability. In order to do this, a series of analyses has been performed. These analyses have entailed computing the correlations and, where appropriate, comparing means between different external measures of students' oral ability and the mean overall test grades awarded by the raters. The two external estimates used were teachers' estimates of students' oral skills and students' self-scores in speaking, calculated from results from the EVA self-assessment questionnaire (see Appendix H on page 290). The rather indirect process of arriving at these measures is described in 'The external aspect of validity' pages 84 – 85. The methods have been designed in order to facilitate the process for the teachers. The measures are clearly only approximate and were not available for every student.

As a precaution, to check the extent to which the test is measuring speaking ability and not something else, correlations have also been found between the speaking test overall grades and the measures available for writing ability (teachers' estimates) and listening and reading ability (test scores).

The Pearson correlation coefficient, r, between the EVA speaking test grades and teachers' estimates, based on 39 students, is 0.53, significant at the $p<0.001$ level). This statistic is regarded as encouraging, considering that the external measure is only an estimate. Moreover, a high correlation would be unlikely, given that teachers are not assumed to be competent in rating oral ability at such an early stage in their involvement in the EVA Project; part of the point of the project was to enhance this competence. The means for the test grades and teachers' estimates are extremely close, being 3.90 and 3.91 respectively, with standard deviations of 1.0 and 1.2.

In the ELTS (English Language Testing Service) validation report (Criper and Davies 1988: 56), the correlation between the ELTS speaking test scores and language tutors' assessment for a sample of 161 students is reported to

have yielded a much lower coefficient of 0.42. This was, in fact, the highest of the test-score/tutor assessment correlations for all the ELTS test components. In an evaluation reported by Wall *et al.* (1994) of the IELE (Lancaster University's Institute of English Language Education) Placement Test, correlations of sub-tests (which did not include speaking) with language teachers' assessments ranged from .78 in the case of grammar (which was exceptionally high) to .47 for writing and reading. Alderson *et al.* (1995) state that 'most concurrent validity coefficients range from .5 to .7 – higher coefficients are possible for closely related and reliable tests, but unlikely for measures like self-assessments or teacher assessments' (1995: 178).

Results from students' self-scores for speaking have also been correlated with the EVA speaking test grades. This correlation, based on 43 students, yielded $r = 0.35$, which is significant at $p<0.01$, and is lower than that obtained in the corresponding ELTS correlation, which was 0.47 (and which, incidentally, was considered very high compared with the self-scores on the other ELTS components). It is worth adding, however, that the overall ELTS test score had a correlation with an overall proficiency self-score of 0.39, which is commented on as 'an unexceptional figure in terms of similar findings elsewhere' (Criper and Davies 1988: 52). In the IELE validation, correlations with self-assessments ranged from 0.51 for listening to 0.3 for writing. The conclusion must be that correlations between self-score and test grade of around 0.3 to 0.4 are to be expected with untrained self-assessors, although higher correlations have been reported where training has been given, according to Clapham (1988: 51), who questions whether there are, in fact, grounds for using (untrained) students' self-assessments at all to validate tests, while not disputing their pedagogic value.

The results are summarised in Table 5.1, which shows the correlations between speaking test grades and external criteria in the EVA and ELTS speaking tests and the IELE writing test (which is the IELE sub-test most comparable with the EVA speaking test, in that it is most direct and is rated across a series of band scales).

Table 5.1 Correlations between speaking test overall grades and external criteria in the EVA and ELTS speaking tests and the IELE writing test

test	**correlation with teachers'/ tutors' estimates**		**correlation with self-score**	
	n	r	n	r
EVA (speaking)	39	0.53	43	0.35
ELTS (speaking)	161	0.42	151	0.47
IELE (writing)	49	0.47	53	0.30

The wider implications of the above findings are disconcerting, and suggest a considerable need for research into the disparities between teacher- and self-estimates and test results, particularly in the present climate of placing assessment increasingly in the hands of learners themselves and of observers of their performance. Bearing this point in mind the findings do at least, suggest that the EVA speaking test compares favourably with both ELTS and IELE in terms of validation against external criteria, especially in the case of teachers' estimates, which appears to be the more worthwhile criterion to measure against. This result is in line with what might have been expected. Teachers are not experts in assessing the spoken language, but have, in most cases in the EVA testing situation, taught the class for almost two years and thus have a shrewd idea of the level of their students' ability. Their estimates were therefore collected with test validation in mind; it would have been surprising if a direct test of this type was either out of line with or identical to these estimates of ability. Students, however, are quite unused to assessing their own ability, and their self-assessment was carried out largely to introduce them to this practice, with washback primarily in mind; it was not expected that this would be particularly useful in providing a means of externally validating the test.

When the overall speaking grades were correlated with measures of other language abilities, i.e. writing, listening and reading, the r-values yielded were still significant (i.e. indicating, unsurprisingly, some relationship between speaking and these other abilities). However, these values were found to be distinctly lower than for the teachers' estimates of speaking ability. This was not the case in the ELTS test, where the correlations of the speaking test with reading and listening were the same as in the tutors' estimates of speaking, and the correlation with writing was, in fact, higher (Criper and Davies 1988: 42). In other words, while there was no evidence from the ELTS correlations that the speaking test was testing speaking as opposed to other skills, a different picture has been presented by the EVA correlations. Here the speaking test scores show convergence with the other speaking measure and divergence with other skill measures, which can only be regarded as an encouraging result. Both sets of correlations are shown in Table 5.2.

Table 5.2 Correlations between speaking test overall grades and other measures of the four skills for EVA and ELTS

	EVA speaking test	ELTS speaking test
writing	0.36	0.46
listening	0.40	0.41
reading	0.41	0.42
speaking (teacher estimate)	0.53	0.42

Notwithstanding the relatively satisfactory result of this validation, there is clearly room for improvement. Moreover, an analysis of the kind just described can only yield results that are as good as its data. In the case of this analysis, not only were the external measures only approximate, but the test grade data itself may also have been flawed. Possible causes of flawed data will be dealt with in the following sections.

The CONTENT aspect: test bias with respect to gender

In 'The content aspect of validity' pages 66 – 71, it was concluded that test bias with respect to gender was a possible threat to the content aspect of validity, which could be investigated through scoring data.

Boys' overall test grades were, on average, slightly lower than girls' (with means of 3.86 and 4.03 respectively). Although efforts were made at the test-development stage (through careful selection of topics, tasks and pictures) to prevent the test from being biased in favour of either boys or girls, the results suggest, on the face of it, that girls have in fact been favoured. However, the possibility that girls were simply better than boys prompted the investigation of the relationship between boys' and girls' mean values of both test grades and teachers' estimates, by running t-tests (using the larger NSD stat dataset). The size of the t-value indicates whether or not the two gender groups can be considered to have been performing differently. In the case of the test grades, data is available for 70 students, while teachers' estimates are available for 315 students. The results are shown in Table 5.3.

Table 5.3 Result from t-testing of teachers' estimates and grades, distributed across gender

	teachers' estimate	**test grade**
n	315 (157 boys, 158 girls)	70 (36 boys, 34 girls)
t-value	-2.19	-0.77
significance ($p<.05$)	yes	no
girls' mean score	4.09	4.03

The t-values in Table 5.3 indicate that the teachers' estimates varied significantly between boys and girls, and that the test scores did not. However, a closer inspection of the mean scores for the two gender groups reveals that, in fact, the test scores and teachers' estimates were very similar, and it is possibly the larger size of the population that caused a significant t-value to be yielded in the case of teachers' estimates. In fact the conclusion must be, on

the basis of this analysis, that the EVA speaking-test scores simply reflect the tendency already indicated in teachers' estimates of speaking ability, i.e. that girls at this stage do slightly outperform boys. This tendency is also reflected in the EVA writing, reading and listening test scores. These results combine to reinforce Bachman's (1990) point, cited in 'Test bias' pages 22 – 23, that group difference in performance may not necessarily be a result of test bias but rather an indication of a difference in the actual language ability of the particular group (1990: 278). By simply reinforcing the impression that the girls' language ability overall was better than the boys', the test appears not to be biased with respect to gender.

Generalisability

Generalisability, as it is studied here, can be regarded as equivalent to reliability. As suggested in Table 4.5 on page 94, the types of potential threat to this aspect of validity that need to be examined are:

- lack of rater reliability

which in turn can be ascribed to

- vagueness of wording in statements in the scoring instruments.

Correlations have been carried out to examine inter-rater reliability. Although this is increasingly investigated by means of multi-faceted Rasch analysis, this has not been attempted here, since, in the present dataset, sufficient information on factors affecting variance in ratings is not available (see 'Data and methods' pages 96 – 99).

Inter-rater reliability

Studying the inter-rater reliability of the test is valuable not only for establishing the consistency, or generalisability, of a test grade, but also for shedding light on which 'sub-skills' seem to influence raters in their grading, and how consistently raters appear to judge these sub-skills. Ideally, intra-rater reliability should also be studied, but the limitations of the dataset do not permit this. This is not regarded as a major drawback, as the purpose is not to test actual individuals (since the test is to be used by an unlimited group of teachers) but rather to see whether different individuals are likely to interpret statements about ability in the same way. In order to fulfil this purpose in the simplest way, only the 16 statements on the performance profile relating to particular *sub-skills* were included in the present analysis, those six relating to *task achievement* having been eliminated. Statements relating to sub-skills are marked with an asterisk in Appendix B pages 274 – 276.

The ratings in the dataset were taken from five raters. All were teachers, or teacher trainers: three were native speakers of English and two were

Norwegians; two were women and three were men. Thus the group can be considered to be typical of the raters who will normally be working with the tests (regarding professional background) and balanced and heterogeneous (regarding native-speaker status and gender). Three raters had been involved in the development and trialling of material from the start, while the remaining two were drafted onto the team shortly before this rating took place.

The findings on inter-rater reliability are considered in three parts, each using the Pearson product-moment correlation coefficient, r. Overall grades are considered first in order to establish whether raters were generally in agreement in setting levels of performance. Scores on the sub-skills are then examined in order to see how true, or generalisable, a profile of ability the test seems to have yielded. Finally, correlations between sub-score skills and overall grades are considered to see which sub-skills appear to have been most influential in setting grades, and what implications can be drawn from this for the reliability of the scoring instruments.

Since raters 2 to 5 generally rated non-overlapping test performances it was not possible to obtain inter-rater correlations between these four raters. Consequently, all raters had to be correlated solely against rater 1 (myself), who had acted as a second rater for all the performances. This is clearly unfortunate, and undermines the validity of the investigation. However, the correlations on overall grades indicate that rater 1 and three of the other raters were, on the whole, in line with each other. Thus the results are regarded as worth presenting here, and are able to give some credible information on which sub-skills were able to be rated with some agreement. Raters 2 to 5 each assessed about 12 tasks which were also rated by rater 1, and consequently the figures used in this section are based on between 12 and 14 ratings from each rater.

OVERALL GRADES

The inter-rater correlations on overall grades are shown in Table 5.4.

Table 5.4 Inter-rater correlations between the overall grades of four external raters and those of rater 1

rater	**2**	**3**	**4**	**5**
	0.40	0.89	0.69	0.84

As can be seen from Table 5.4, the correlations range from 0.4 to 0.89. However, most raters have a correlation ranging from approximately 0.7 to 0.9 (significant at $p<.01$) with rater 1, while rater 2 has a low correlation of 0.4 (not significant). In order to evaluate these correlations, it is necessary to

consult other, comparable studies. In the case of ELTS validation (Criper and Davies 1988), the inter-rater correlations for the speaking test are not released, but the figures for the writing test range from 0.54 to 0.78, with most being around 0.64 (1988: 105). The authors maintain that, in order to claim to have an accurate measure (of writing), the 'first criterion must be a high level of inter-rater reliability, around 0.9 (a reliability claimed, for example, by the ETS Test of Written English)' (1988: 105).

Whether such a demand can realistically be made of a test of speaking is questionable. Shohamy *et al.* (1986) cite inter-rater correlation coefficients on four separate oral tasks as 0.91, 0.81, 0.76 and 0.73, and claim 'It is clear from these analyses that the reliability of the tests was relatively high' (1986: 216). Douglas (1994) describes the coefficients on four parts of his oral test – 0.78, 0.76, 0.78 and 0.77 – as suggesting 'a fair degree of inter-rater reliability' (1994: 127).

The degree of inter-rater reliability on overall grades in the EVA speaking test, as far as coordination between the main rater and three out of the four others is concerned, can be considered comparable with the studies cited, although not high. In the case of the other rater it is clearly unacceptable, at 0.4. In retrospect, this is an unsurprising result, given that the rater who was out of line had not worked with the team for any length of time, and the need for rater training, even in the case of those who had worked together on the test design is highlighted. Investigation at a deeper level into other possible causes of poor inter-rater correlation is clearly necessary.

SUB-SKILL SCORES

In order to shed light on what aspects of performance raters were able to agree on, inter-rater correlations were next obtained for each of the 16 sub-skills in the performance profile. Table 5.5 shows the r-values for the four raters with respect to rater 1.

Sub-skills are placed in four groups, generally according to how many raters have r-values greater than 0.6, ranging from all raters in group 1 to one rater in groups 3 and 4 (with the latter group containing cases where two correlations are very low). Rater 2's correlations are shown in the last column, but have not been included for ranking purposes as this rater's overall grade inter-rater correlation has been found to be so low. It is more interesting to see how raters varied in their sub-skill judgements when they were fairly closely in line in rating overall level.

A first point to be made is that the inter-rater correlations for the individual sub-skills in each column are generally lower than for the overall grade shown at the bottom. This is line with what Alderson (1991: 80) has observed about sub-scores and global grades.

Table 5.5 Correlation coefficients for four raters with respect to rater 1: sub-skills

group	sub-skill	rater 3	rater 4	rater 5	*rater 2*
1	language idiomaticity and independence	0.82	0.72	0.71	*0.75*
	reading pronunciation	0.65	0.73	0.84	*0.79*
	reading flow	0.78	0.69	0.71	*0.57*
	initiative	0.80	0.74	1	*0.46*
1/2	strategies	0.66	–	0.65	*0.89*
	style and politeness	0.70	0.69	–	*0.44*
2	vocabulary	0.72	0.78	0.57	*0.39*
	language correctness	1	0.31	0.79	*0.09*
	intonation (suggesting friendliness)	0.76	0.37	0.70	*0.30*
	intonation (carrying message in free speech)	0.62	-0.35	0.76	*-0.13*
3	reading intonation	0.80	0.56	0.48	*0.51*
	keeping going	0.59	0.36	0.79	*0.47*
	contribution	0.44	0.41	0.84	*0.53*
4	pronunciation (free speech)	0.09	0.61	0.38	*0.46*
	special expressions	0.24	0.69	0.32	*0.52*
	sounding friendly and interested	0.89	-0.01	-0.5	*0.12*
	overall grade	0.89	0.69	0.84	*0.40*

Considering the sub-skills related to sounds, it is interesting to note that both *pronunciation* and *flow* in reading are highly correlated, while pronunciation generally (i.e. in free speech) has very poor correlation. In the case of *intonation*, there is little difference between the degrees of correlation in reading and in free speech.

High correlations are obtained for core 'language' sub-skills, such as *language idiomaticity and independence*, and, to a lesser extent, *vocabulary* and *correctness*. *Strategies* and *style and politeness* seem to be fairly highly correlated, although the dataset is incomplete for these sub-skills. Of the sub-skills most associated with 'message and fluency', *initiative* is highly correlated, while in the case of *keeping going* and *contribution* the correlations are quite low.

Various explanations can be offered for the differences in degree of correlation. In the case of phonological features, it seems reasonable that sounder, more focused judgements are made while listening to a passage of reading. In the case of the more core language skills, such as *vocabulary* and *correctness,* teachers are probably more used to considering these in their assessment of language ability (including written) generally.

A scrutiny of the wording in the statements where correlations are low suggests that the wording employed in these may have been ambiguous. Keeping going, contributing and sounding friendly, in the absence of any more concrete explanation, can mean different things to different raters. On the other hand, more tangible notions, such as the degree of initiative-taking or the absence of vocabulary gaps, are probably easier to recognise with consensus. And the statements relating to strategies and style/politeness appear to be quite clearly explained and exemplified on the profile form. In order to find out how far these hunches are correct, a further investigation has been carried out on raters' own perceptions of vagueness or ambiguity in the wording of descriptors. This analysis is reported on in 'Vagueness in the wording of the scoring instruments' pages 110 – 112.

At this stage, it is not possible to conclude whether it is the actual wording of the questions on the profile, or the effectiveness of the task in eliciting evidence of abilities, or rater familiarity with the sub-skill itself, which has the greatest influence on the correlations shown here. However, it seems that certain loose hypotheses are indicated from these findings about inter-rater reliability, and can be summed up as follows:

- in general, there is greater inter-reliability on overall grades than on sub-skill scores
- there is greater inter-reliability on the more core 'language' sub-skills, such as *idiomaticity* and *vocabulary* than on sub-skills related to 'fluency and message', such as *keeping going* and *contributing*
- there is greater inter-reliability in judging sub-skills with definite, tangible wording than with vague, abstract wording
- there is greater inter-reliability in judging phonological features in reading than in free speech.

The question of how far the vagueness of the terminology in the rating instruments was actually a threat to reliability remains to be addressed. However, it is a fact that out of the 16 sub-skills, only four are directly related to message and fluency – *keeping going, contribution, initiative* and *strategies* – and of these only the last two have high inter-rater correlations. This gives good reason to believe that the assessment based on the performance profile provides a less reliable basis for choosing a level on the fluency-related scale than on the language-related scale. In the latter scale, 12 sub-skills are involved to some degree, and the most core of these (such as *idiomaticity, correctness* and *vocabulary*) show relatively high levels of reliability.

THE APPARENT INFLUENCE OF SUB-SKILL SCORES ON OVERALL GRADES

If the raters' instructions have been followed, the overall grade should have been influenced at least as much by the setting on the fluency-related scale as by that on the language-related scale. This would imply that scores on the sub-

skills related to fluency should have considerable influence on the setting of the overall grade. In order to investigate whether this has proved to be the case, correlations between overall grades and sub-skill scores have been calculated for each rater. By comparing the findings of this analysis with that of the previous one, implications can be made about the reliability of the overall grades. If these appear to have been heavily influenced by sub-skills that were difficult to agree on, this will detract from their potential reliability. In this analysis, the number of sub-skills has been reduced to 13, as the phonological features tested in free speech, as opposed to reading, have been excluded. Table 5.6 illustrates the correlations between sub-skill scores and overall grades, ranked according to the correlations of the mean sub-skill scores with the mean grade.

Table 5.6 Correlations between sub-skill scores and global grades for each of the five raters

sub-skill	Mean score/ grade	rater 1	rater 2	rater 3	rater 4	rater 5
contribution	0.78	0.68	0.65	0.85	0.57	0.85
vocabulary	0.76	0.84	0.60	0.90	0.49	0.65
language idiomaticity and independence	0.74	0.78	0.70	0.90	0.67	0.85
language correctness	0.74	0.76	0.64	0.91	0.54	0.90
reading flow	0.66	0.67	-0.70	0.80	0.48	0.76
keeping going	0.65	0.66	0.75	0.94	0.83	0.90
reading intonation	0.64	0.59	0.46	0.84	0.72	0.87
strategies	0.63	0.73	0.74	0.72	0.39	0.85
reading pronunciation	0.61	0.57	0.40	0.63	0.61	0.62
initiative	0.59	0.64	0.89	0.96	0.86	0.80
special expressions	0.55	0.61	0.30	0.60	0.55	0.55
style and politeness	0.53	0.49	-0.01	0.54	0.56	0.49
sounding friendly and interested	0.47	0.42	0.24	0.74	0.39	0.58

The table shows that the more core language-related sub-skills, *vocabulary, language idiomaticity and independence* and *language correctness*, as well as those related to message and fluency, *contribution, keeping going, strategies* and (rather less so) *initiative*, have correlations between mean scores and global grades in the region of 0.6 to 0.8. This suggests that these sub-skills have been influential in the setting of global grades. These findings were to be expected, since the overall grade was given on the basis of the setting on two band scales in which these two groups of sub-skills featured fairly explicitly. Interestingly, those features that seem least influential, i.e. where the correlation of mean scores and global grade were clearly less than 0.6, were those that might be associated with pragmatic ability – *special expressions, style and politeness* and *sounding friendly and interested.*

The findings here, coupled with those of the previous section, bear out some significant conclusions. The most core linguistic sub-skills, as exemplified above, have had a clear influence on grade. Given the fact that these sub-skills were generally rated consistently across raters, this must be seen as a strength of the test.

However, as can be seen from Table 5.5 on page 107, two of the four sub-skills in the fluency-related group – *keeping going* and *contribution* – although seemingly influential in raters' grade-setting, were relatively inconsistently assessed across raters. This must be seen as a contributory factor to weakness throughout the rating procedure, from the filling in of the profile performance to the setting of the overall grade. Clearly, raters must be given more help in rating these qualities, for example by wording the relevant questions on these sub-skills in more concrete terms.

Vagueness in the wording of the scoring instruments

Vagueness in the wording of the scoring instruments has been cited as a source of potential weakness in the test, particularly in the case of the elements of ability related to fluency. In order to attempt to pinpoint where in the scoring instruments the wording can be regarded as vague, or difficult to interpret, an investigation has been carried out not using scoring data, but using raters' opinions, collected in retrospect, on the various parts of the instruments. Raters were asked to assign numbers to each broad band, or level, on the two band scales *language structures and vocabulary* (L/V) and *message and fluency* (M/F), shown in Appendix C pages 277 – 279, as well to as each sub-skill statement on the performance profile, shown in Appendix B pages 274 – 276. It must be emphasised that the wording being judged is that in the actual statements in the band scales and profile form, and not on the abbreviated denotations for the sub-skills used in tables and graphs here. The sub-skills involved were the 13 listed in Table 5.6 on page 109. Raters were asked to assign points as follows:

1 = very clear and unambiguous

2 = on the whole, clear and unambiguous

3 = rather vague and open to different interpretations

4 = very vague and open to different interpretations

Four out of the original five raters supplied this information and the results of the survey are shown in Tables 5.7 and 5.8.

The results of the survey are at the same time reassuring and disappointing. In the case of the band scales, as shown in Table 5.7, there was a general consensus that these were on the whole clear, with a mean of less than 2, but that the middle band in each (i.e. grades 3–4) was least clear, with mean values

Table 5.7 Points awarded on 'vagueness' of wording on three levels in each of the band scales, M/F and L/V, judged by four raters

band scale	level	mean vagueness score
message and fluency (M/F)	top (grades 5–6)	1.75
	middle (grades 3–4)	2.5
	lower (grades 1–2)	1.75
language structures and vocabulary (L/V)	top (grades 5–6)	1.75
	middle (grades 3–4)	2.25
	lower (grades 1–2)	1.75

Table 5.8 Sub-skills ranked in order of clarity as perceived by raters

group	sub-skill	mean points
1	reading pronunciation	1.33
2	special expressions	1.66
	style and politeness	1.66
3	initiative	2.33
	contribution	2.33
	vocabulary	2.33
4	keeping going	2.66
	reading flow	2.66
	reading intonation	2.66
	language correctness	2.66
5	language idiomaticity and independence	3.0
	strategies	3.0
	sounding friendly and interested	3.0

in excess of 2; this was particularly the case for the M/F scale. One explanation for this might be the fact that the middle bands on both scales contained considerably more text than the upper and lower.

In the case of individual sub-skills referred to in the performance profile, the result was rather muddled, since raters differed very widely in their opinions, spanning a range of three points in four of the thirteen sub-skills. Sub-skills are ranked in order of clarity, and placed in five groups, according to the mean number of points awarded by raters, in Table 5.8.

When comparing Table 5.8 with Table 5.5 on page 107, which shows inter-rater correlation coefficients, no clear correspondence of clarity–reliability can be found. *Special expressions*, regarded as one of the most clear and unambiguous, had an extremely low inter-rater correlation coefficient. On the other hand, two of the elements generally regarded as most vague were among

those with the highest inter-rater coefficients (*language idiomaticity and independence* and *strategies*). While there were other cases, such as *reading pronunciation* where the expected correspondence was found, the hypothesis proposed in 'Vagueness in the wording of the scoring instruments' pages 110 – 112 – that a straightforward relationship exists between inter-rater reliability and perceived clarity in the wording of descriptors – is not generally supported by the evidence here.

The fact that concepts more closely related to having language resources, such as *vocabulary, pronunciation* and *idiomaticity*, seem to have been rated more reliably than those related to using these resources fluently, e.g. in *keeping going* or *contribution,* may be explained by other factors than the wording itself. It could be that raters relied more heavily on their own personal, idiosyncratic notions of what goes into fluent performance, and that what might have seemed an obvious interpretation of a fluency descriptor to one rater might have been far from obvious to another. For example, *keeping going* can clearly mean two quite different things to two different raters. This explanation is in line with those of researchers such as Esser (1996) and Freed (1995), and is further discussed in 'Fluency' pages 124 – 135.

At the same time, it must be emphasised that vagueness in the wording of descriptors is always undesirable (see North 1997) and, in itself, can only detract from the reliability of a scoring instrument. The demand for clarity and concreteness in the wording of descriptors is in no way diminished by the findings here. What has emerged is an indication that other factors, such as a lack of universal consensus on the interpretation of fluency and its related concepts, such as *contributing* and *keeping going*, may have had a greater effect on reliability than the actual wording of descriptors.

Conclusions on generalisability

The conclusions to be drawn from the investigation into the generalisability of the scores and grades of the EVA speaking test largely bear out what was expected in terms of strengths and weaknesses of the test. The overall test grades for most raters were fairly satisfactorily correlated with the central rater, although in the case of one rater this was not the case. However, given that the final grades were worked out as the mean of the two ratings, one can assume that these grades represent fair indications of spoken performance generally, at least for three quarters of the group. This can also be said of many of the sub-skill scores on the profile form, particularly where the more core linguistic sub-skills are concerned. Because of this strength, the setting on the *language structures and vocabulary* band scale, which is made largely on the basis of these sub-skill scores, can probably be regarded as reliable, particularly when rater training is employed, as is the intention.

However, the low inter-rater correlations on the sub-skills associated with message and fluency is a matter of greater concern. Because this group of sub-

skills are apparently influential in the setting of the overall grade, they can be assumed to reduce the reliability of this overall grade, and hence the setting on the *message and fluency* band scale. This finding was anticipated in 'The generalisability aspect of validity' pages 83 – 84, and was initially attributed to the fact that these sub-skills seemed generally to be described in rather vague terms, and hence were open to multiple interpretations; this has, however, not been supported by the judgements of raters. The question has, therefore, arisen as to which other factors might be involved in reducing the inter-rater reliability of elements related to fluency. It has been suggested that lack of inter-rater agreement may be inherent to fluency-related aspects of performance, which underscores the need for a more specific way of describing these, based on what is unambiguously observable in performance. The fact that only four of the sub-skills assessed are fluency-related is seen as a further detractor from the reliability of the scoring on this side of performance, specifically through the M/F band scale. (Since there are several questions on the profile form, not analysed here, which relate to achievement on specific tasks, 'message' as such cannot be regarded as under-represented, and the deficiency is thus regarded as lying in the representation of 'fluency'.)

Thus, it seems that the generalisability of the test results is dependent on systematic use of rater training as well as on an increase in the number of sub-skills relating to fluency, and on finding more specific, unambiguous, data-informed ways of assessing fluency.

The STRUCTURAL aspect

For band scales to work, it is not enough to ensure that they are reliable, i.e. that different raters set the same performance at more or less the same levels when using the scales. So that they can be used differentially to make true statements about the various sides of language ability, the scales must reflect a true picture of how this ability is made up, in the way they are structured. Each scale should represent an area of ability composed of related elements, such that the level of performance on one of these elements should more or less predict the level of performance on another. On the other hand, performance rated on one scale should not predict performance rated on another (otherwise there would be little point in having separate scales). In 'A validation framework' pages 30 – 31, a potential area of threat to the structural aspect of validity was defined as the lack of primary and secondary evidence to support the way elements, or constructs, of ability are clustered and divided in the band scales. In 'The validation process' pages 65 – 87, secondary evidence, mainly in the form of established testing procedures, was generally found to support the division of constructs into language-related on one scale, and message/fluency-related on the other. It remained, therefore, as indicated in Table 4.5 on page 94, to examine the extent to which primary evidence based on scoring data might give

further support to this clustering and division of constructs.

In order to investigate this, factor analysis has been performed on the scores (averaged across raters) for 13 ability-related sub-skills (i.e. excluding task-achievement sub-skills) on the performance profile. Factor analysis is able to identify broad areas of underlying ability, or factors that account for the way students' scores vary in a test. Within each of these factors, a score on one sub-skill will roughly predict the score on another; the scores are said to share common variance. Across the factors, no such prediction can be made.

In the case of the EVA test, factor analysis was able to reduce the possible 13 different scores for each student to a more manageable number of scores on factors, each of which could, ideally, be identifiable as relating to a particular aspect of ability, e.g. pronunciation. However, the point of using factor analysis here was not actually to produce a number of different scores, but rather to see how the sub-skills seemed to cluster together, and what kind of factors – or broad areas of underlying ability – would thus emerge as separately assessable. It was hoped that this would justify the way abilities have been grouped together in the EVA test for the purposes of reporting on a student's ability, i.e. through the use of separate L/V and M/F band scales, and with pronunciation/intonation assessed as an adjuster.

The first solution produced in the factor analysis was the 'orthogonal solution', shown in Table 5.9. In an orthogonal solution, factors share no common variance with each other; i.e. a student's ability on one factor is apparently unrelated to that on the others.

As is typical for orthogonal solutions, Factor 1 is some kind of general ability factor, shown by the fact that all the sub-skills load highly on this

Table 5.9 Orthogonal solution from factor analysis of mean

mean sub-skill score	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 4**
vocabulary	.817	-.041	-.145	-.244
reading pronunciation	.778	.292	-. 8	.231
reading flow	.79	.182	.278	.027
reading intonation	.681	.279	-.265	.506
style and politeness	.712	.162	.461	.168
sounding friendly and interested	.639	.281	.5	-.163
special expressions	.723	.133	-.08	-.43
strategies	.789	.184	-.242	-.16
keeping going	.774	-.461	.201	.015
initiative	.68	-.591	.094	.111
contribution	.835	-.41	-.049	.106
language correctness	.826	.021	-.39	.033
language idiomaticity and independence	.875	.021	-.193	-.146

sub-skill scores

factor. The value of the loading can be seen as indicating the degree of correlation between the sub-skill scores and a score hypothetically worked out on the whole factor. Following Child (1970), a loading whose absolute value is greater than 0.3 is regarded here as salient (1970: 39).

On Factor 2, keeping going, *initiative* and *contribution* have salient loadings. On Factor 3, *style and politeness* and *sounding friendly and interested* have salient loadings (as well as *language correctness*, although with a different polarity). On Factor 4, *reading intonation* and *special expressions* have salient loadings, with opposite polarities. Thus it is already apparent that a pattern is emerging that allows some speculation about the way language ability is composed. However, the orthogonal solution gives an imbalanced picture, as almost all of the variance in scores (74 per cent) is accounted for by Factor 1, with none of the other factors accounting for more than ten per cent of the total variance.

In order to get a more balanced picture of the way different underlying areas of language ability seem to account for the variance in test scores, an oblique solution was sought. This solution redefines the factors somewhat so that no single factor dominates, each factor contributing significantly (in the present case between 17 and 32 per cent) to the common variance in the test scores. The disadvantage of the oblique solution is that the factors are no longer independent, sharing some variance with each other. This means that a score on one factor would, to some extent, predict a score on the other factors. However, this joint variance is small, particularly in the case of Factors 1 to 3, where it is less than four per cent. The oblique solution (primary pattern) is presented in Table 5.10.

Table 5.10 Oblique solution from factor analysis of mean

mean sub-skill score	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 4**
vocabulary	.532	.401	.014	.007
reading pronunciation	.105	-.151	.37	.613
reading flow	.033	.03	.721	.131
reading intonation	-.006	-.179	-.043	1.07
style and politeness	-.285	.023	.935	.132
sounding friendly and interested	.075	-.177	1.064	-.224
special expressions	.745	.102	.197	-.179
strategies	.637	.039	.002	.281
keeping going	-.247	1.021	.214	-.171
initiative	-.354	1.181	-.049	-.08
contribution	-.102	.964	-.096	.147
language correctness	.474	.299	-.308	.525
language idiomaticity and independence	.515	.325	-.006	.198

sub-skill scores

Here Factor 1 is no longer a general ability factor, but is significantly loaded on by a distinct group of sub-skills: *vocabulary, special expressions, strategies, language correctness* and *idiom/independence*. The way sub-skills load significantly on factors is shown in Table 5.11.

Table 5.11 The way subskills load significantly on factors, using the oblique solution

Factor 1	*Factor 2*	*Factor 3*	*Factor 4*
vocabulary	keeping going	style and politeness	reading intonation
special expressions	taking initiative	sounding friendly and interested	reading pronunciation reading pronunciation
strategies	contribution	reading flow	language correctness
language correctness	vocabulary*	reading pronunciation	
language idiomaticity and independence			

* to a much smaller degree

The nature of the sub-skills that load saliently on factors gives an indication of the 'flavour' of each factor. Factor 1 is thus associated with more core linguistic abilities (non phonological), while Factor 2 is mainly associated with elements relating to message and fluency. Factor 3 is primarily associated with what can probably be regarded as personality factors of politeness and 'niceness', while Factor 4 is clearly associated with pronunciation and intonation, but also with language correctness (which can of course be judged on pronunciation and intonation).

These associations are not watertight. For instance, the sub-skill *strategies*, as it is represented in the profile (i.e. the non-use of L1), loads on the core language factor, not the message/fluency factor, as might have been hoped, seeing that this sub-skill was intended to be placed on the M/F scale. However, it would be premature to reconsider this decision on the basis of this one analysis. It is possible, for instance, that the wording of the descriptor on the use of strategies might have been responsible for leading to a more language-focused assessment.

In fact, excessive credence should not be placed in this evidence generally. As was the case of the inter-rater reliability coefficients, the low number (four) of message/fluency related elements detracts from the implications that can be drawn from the result of an analysis based on the present set of sub-skill scores. Clearly, a future analysis should be performed using ratings collected

through a revised set of band scales and performance profiles. However, these preliminary results must be regarded as lending support to the convention of using separate band scales to rate language and fluency, and to the practice, observed in the EVA test, of rating pronunciation and intonation independently of other abilities.

Summary

In this chapter, rating data has been used to see how far the findings – or unanswered questions – of the a priori investigation in Chapter 4 can be supported, or illuminated, by statistical evidence. The results have produced no major surprises, but have reinforced some important conclusions on the validity of the EVA speaking test. A reappraised profile of the test's validity, showing its strengths and weaknesses (in terms of need for improvement), as well as areas that remain in need of further investigation, is presented in Table 5.12 on page 119.

The external aspect of validity based on teachers' estimates has been found to be comparable with other, similar, validation studies, although it was not found to be satisfactory in the case of convergence with other measures of speaking. The problems may lie at least as much with the other measures – teacher estimates and self-assessment – as with the test itself, and the need for research into alternative ways of assessing speaking ability is highlighted. The data did however show encouraging divergence with measures of other skills. No evidence was found to suggest that the content aspect was impaired through gender bias. The generalisability aspect could be improved significantly by enhanced inter-rater reliability. This can be brought about to some extent by ensuring that comprehensive training is given to raters. The most significant finding regarding rater reliability, however, is that certain sub-skills, although influential in the setting of the overall grade, have been judged very differently by the individual raters. This particularly concerns fluency-related sub-skills, where it is recommended that more sub-skills should be assessed on the performance profile, and that these should be worded in a less ambiguous way. For the rating procedure to work in the way intended, this recommendation should be extended to cover the band scale for *message and fluency*. It has been found that, on the whole, core sub-skills relating to the *language structures and vocabulary* band scale were more reliably rated, and that, therefore, settings on this band scale, and scores on the associated sections of the performance profile, appear to be potentially generalisable. A boost has been given to the structural aspect of validity, through the factor analysis of sub-skill scores. The result of this analysis indicates that there is justification for the practice of assessing the language and fluency aspects of performance in different band scales and for giving a

separate assessment of pronunciation and intonation.

While it has not been considered appropriate to use rating data in the empirical investigation of consequential aspects of validity (since this can only be studied over a period of time following the introduction of the test), certain of the conclusions drawn above must be viewed with consequential aspects of validity, and specifically washback, in mind. This largely concerns the recommendation that more sub-skills should be assessed, and that these should be worded in a more concrete way. In the interests of the consequential aspect of validity, two important caveats must be added to this recommendation. The first is that any statements built into the rating instruments should be predominantly positive, i.e. 'can do' rather than 'cannot do'. This is vital in giving students encouragement, even at low levels, as well as in giving them actual aims to work towards. The second is that statements should, as far as possible, concern actual language use, and hence have didactic value. Although there is a case, in the interests of reliability, for referring to such recognisable 'symptoms' as excessive disruptive hesitation and rate of speech, these have little didactic value to a student who needs to know how to improve his/her fluency. It is therefore essential to include statements that refer to the use of actual linguistic devices for bringing about fluency. It must, moreover, be emphasised that the conclusions reached regarding the washback effect of the scoring instruments should not be restricted to those used by teachers, but should be incorporated into a set of instruments for self-assessment by students.

While the implementation of the recommendations made here, and in 'Conclusions on the validity test' pages 90 – 95, requires a long-term period of adjustment of the test, certain questions remain to be answered before this can be embarked on. The most fundamental of these involves the problematic matter of defining unambiguous statements concerning actual language use that relate to student fluency at different levels of ability. The work of defining more aspects of fluency, in more concrete ways, using the data of students' language, as is to be done here, provides a response to calls for data-driven rating scales of fluency, such as that made by Fulcher (1996). However, the didactic, linguistic nature of the features of fluency to be investigated in this study poses further challenges. The role of smallwords in meeting this challenge has already been touched on, e.g. in 'The structural aspect of validity' pages 74 – 82. The search for a solution to the questions of what exactly makes up fluency, how it may be identified at different levels, and of the central role of smallwords as recognisable players in bringing about fluency, is taken up as the central theme of the remainder of the book.

Table 5.12 Reappraised profile of the validity of

aspect of validity	judged to be satisfactory	identified needs for improvement	to be investigated in this study	recommended for future investigation
CONTENT	no significant gender bias in tasks tasks designed to elicit language products that contain representative evidence of ability		scoring data findings on gender effect	survey among teachers/students for clarity/familiarity and potntial of tasks to elicit evidence of CLA examine transcripts to monitor test procedures
SUBSTANTIVE	tasks designed to put testees through processes representative of language use			survey among students for perceived authenticity of communication
STRUCTURAL	clustering and division of constructs supported by primary and secondary evidence band scale descriptors supported by some primary evidence	more comprehensive descriptors of performance esp. fluency scale	transcript findings on fluency and role of smallwords	need for more primary evidence to support band-scale descriptors at different levels of ability and language/fluency clustering/division of elements
GENERALISABILITY (subsumes content, substantive, structural and external aspects)	relative representativeness of tasks	training of raters less ambiguous wording in descriptors		scoring/transcript data findings from a larger group on weak/dominant partner effect
EXTERNAL	recognised external criteria used coefficients for divergence (other skills) encouraging	external validation coefficients for convergence (speaking) too low		need for research into alternative ways of assessing speaking ability
CONSEQUENTIAL (subsumes content, substantive and structural aspects)	measures taken to ensure validity, with respect to: link CLA model/curriculum; direct nature of testing; representativeness and authenticity of tasks	positive comprehen-sive, 'linguistic' descriptors self-assessment		need for surveys among test users on use and consequences of testing and appropriacy to new curriculum

Part Two:
Fluency and smallword use

6 Fluency and smallwords – making the connection

In Part One, the EVA speaking test was subjected to a validation process which resulted in a profile of its strengths and weaknesses, the identification of certain needs for improvement, and recommendations for imminent and future research. These findings on the test's validity status are summarised in Table 5.12 on page 119. As the table shows, one investigation is earmarked to be carried out within this study, viz., that of looking into what test transcripts reveal about fluency at different levels, and the part played by smallwords in contributing to fluency. Part Two is entirely devoted to this investigation.

It is worth noting that, while the investigation in Part One centred on a particular test, that in Part Two focuses on spoken fluency in the learner language (English) more generally. Admittedly, the language studied is elicited by the EVA test, but the framework for studying smallword use, as well as the conclusions drawn on its acquisition and consequences for fluency should be of value to anyone involved in teaching/learning the spoken language as well as to those designing means of assessing it.

Without a greater understanding of fluency, and particularly of linguistic features that seem associated with it, it would not be possible to attend to the needs identified in Part One for test improvement. These needs concern the scoring instruments, particularly where they relate to fluency in performance. The call is specifically made for more empirically-founded statements that describe fluent performance at different levels, which are definite and, preferably, positive and linguistic (i.e. referring to actual language use), and which can be adapted for use in self-assessment instruments. Only through the incorporation of such statements can the test be regarded as having validity with regard to the structural, generalisability and consequential aspects, as these were defined in 'Six central aspects of validity' pages 29 – 30. This involves not only looking into the different factors that seem to make up fluency, but also developing some way of measuring these factors so that statements can be made about how fluency manifests itself at different levels.

In 'The structural aspect of validity' pages 74 – 82, a working definition of fluency was derived from Bygate's (1987) claim that speaking involves, over and above an understanding of the system of the language, 'making decisions rapidly, implementing them smoothly, and adjusting our conversation as unexpected problems appear in our path' (1987: 3). Bygate's account of the specific skills required for speaking (e.g. as opposed to writing or reading) was

outlined and expanded in 'Speaking' pages 43 – 46. Five specific skills were listed and then incorporated into the components of CLA, operationalised in Chapter 3 pages 33 – 57:

- skills required to 'play for time'
- skills required to involve or acknowledge the interlocutor, or his/her utterances
- skills required to structure, or 'place', utterances in the discourse
- skills required to cope with potential problems in communication
- skills required to express vagueness and lack of total commitment.

These skills, as they stand, do not contain elements that are readily measurable, so they are not directly able to provide a basis for writing descriptors of performance at different fluency levels. However they are valuable, not only as an underlying explanation of what goes into bringing about fluency, but also in highlighting the scope of the role of smallwords; each of the five skills is shown to be manifested by smallword use.

The problems associated with finding a way of pinning down and recognising fluency have been a thorn in the side of test-makers, researchers into SLA methods, and those involved in the process of learning foreign languages for many decades. The two former groups have been mainly concerned with finding dependable, recognisable markers for measuring fluency – such as rate of speech or the frequency and nature of pauses. The latter group, on the other hand, need to know what creates fluency – such as the acquisition of formulaic language – rather than its symptoms. This chapter begins by laying on the line some of the problems encountered in defining fluency. Next, a summary is given of what appear to have emerged as recognisable markers of fluency, with an account of some attempts to identify elements that seem to discriminate between more and less fluent performance, as well as of linguistic devices whose acquisition seems to enhance fluency. This discussion leads into the proposal that smallwords occupy a significant position among these devices.

Establishing a theoretical basis for explaining just <u>how</u> smallwords can bring about fluency, ultimately leading to a systematic way of analysing the individual functions performed by smallwords in the process, is the subject of the remainder of the chapter. The works of a range of writers with various focuses – either on smallwords themselves or on spoken communication generally – are consulted. Sperber and Wilson's (1995) 'relevance theory' is highlighted as offering the most cohesive explanation for the way smallwords work as a system for effecting fluency. This explanation is reviewed in the light of Levelt's (1989) theory of speech production. The chapter culminates in a framework, founded on relevance theory, which can be used to analyse the signals sent by smallwords in contributing to fluency in the language of students.

The findings of this chapter form the basis for the investigation in the remainder of Part Two. Corpus analysis is employed, comparing the test transcripts of two Norwegian student groups – defined as *more* and *less fluent* – the grouping being primarily based on test scores, but supported by non-linguistic measures of fluency. A corpus of transcripts from a native-speaker control group is also analysed. The initial focus is on revealing the extent to which smallwords are used in nativelike quantities and distributional patterns by the speakers at different fluency levels. The investigation proceeds by considering the nativelikeness of the signals that smallwords are used to send by the two Norwegian groups, and thus presents an account of the effect of use/non-use of smallwords on student language at different levels of fluency. The study concludes by summarising and drawing together the findings from both parts, suggesting ways in which the knowledge acquired in Part Two, concerning fluency and smallword use, might be used to remedy the deficiencies in the test's scoring instruments, as revealed in Part One.

Fluency

Pinning down fluency

Freed (1995) states: 'Just what is meant by the term 'fluency' is rarely if ever discussed. It may be assumed by those who use it – teachers, students, educators, the public at large – that there is some tacitly agreed-upon meaning for the term, but nothing could be further from the truth' (1995: 123). Yet, the concept of fluency is widely referred to, not only in testing, where it tends to be one of the aspects of language assessed, but also in popular statements about language ability, such as 'she speaks fluent French'. These two uses of the term correspond to what Lennon (1990) calls the narrow and broad senses of fluency. The focus in this study is on the narrow, testers' sense, beginning with an attempt to ascertain just how much general agreement there is on what constitutes fluency. We need to know whether the term can be used in scoring instruments with no explanation, as for instance *vocabulary* normally is, or whether it needs to be spelt out, and if so, how this should be done.

Esser (1996) has conducted an investigation into the rating of oral fluency in German among a group of British university students, specifically posing the questions of, firstly, whether raters are consistent (with themselves or other raters) in judging fluency (with no definition offered), secondly, whether they consider the same aspects of performance while rating fluency, and, thirdly, whether raters are in accord when giving definitions of what makes up fluency.

Freed's contention that there is no such thing as a tacitly agreed-upon meaning for fluency was corroborated by Esser's results. While the raters, in carrying out paired-ranking of candidates, tended to be internally consistent, there was very little inter-rater consistency in the ranking. The aspects of

performance that raters claimed to be concerned with while judging fluency covered a wide range, although 'temporal' qualities, e.g. 'flow', pausing and speed, were the most commonly occurring features. Actual references to language quality, e.g. grammar and vocabulary, were made by a number of raters, but were ranked lower than a general reference to comprehensibility. The combination of these findings lead Esser to state that 'we have to treat judges as individuals with their own personal conception of fluency' (1996: 86). When asked to define fluency, only one of eleven raters was able to give any clear definition, and there was considerable variety in what the others offered as features of fluency. Esser's conclusion is that it is unreasonable to expect raters 'to come to a reliable judgement if they have to rate a vague concept with multiple definitions like fluency' (1996: 93).

It seems then, that if fluency is to be assessed at all reliably, it requires a detailed explanation of what it entails. An attempt to capture the essence of this elusive thing is made by Koponen (1995). His extensive review of what has been read into the meaning of fluency is illustrated with a list of 93 attributes. His historical overview of the use of the term reflects the wide diversity of meanings already commented on here. However, the basic ideas inherent to the word 'fluent', such as *flow, unbrokenness*, and *smoothness*, tend to dominate. Recurring, core facets of speaking related to these ideas are typically *rate of speech, lack of excessive pausing, coherence, length of utterances, continuity* and *connectedness*. References to actual areas of language knowledge, such as grammar and vocabulary, do occur, but these elements tend to be peripheral, as in the case of Esser's study. These may perhaps be regarded as significant, not so much in terms of accuracy, which is often considered as somehow 'external' to fluency, but in contributing to what must surely also be regarded as a core facet of fluency: *comprehensibility*.

Whatever core facets might be listed, however, Koponen believes that 'linguistically describable categories and criteria alone are probably not sufficient in themselves, or they may be too coarse, too absolute or abstract or elusive for the analysis of fluency' (1995: 5), which suggests that a listener's perception of fluency is based on an interaction between the spoken performance and his/her own mindset.

The intangible nature of fluency – as a holistic listener-response to a piece of speaking – seems to be one of its salient characteristics. It would be wrong to rob fluency of this characteristic, which we will return to later in this chapter, where a relevance-theory view of fluency is proposed. However, for the purpose of the immediate quest, it is necessary to resort to 'describable categories and criteria' so that fluency can be recognised in performance. This means considering the core facets cited above as being traditionally perceived as elements of fluency, as well as attempting to identify types of language behaviour and elements of actual language use that seem to go hand in hand with fluent speech.

Identifying elements of fluency

The core sense of fluency might be summed up by statements such as Hedge's (1993): 'It is the ability to link units of speech together with facility and without strain or inappropriateness or undue hesitation' (1993: 275). The gradual development of fluency in terms of this ability is encaptured and expanded on in band-scale descriptors typified by that cited by Weir (1993) (see Table 6.1), where fluency is assessed alongside five other components of oral performance.

Table 6.1 Example of criteria of assessment of fluency

0	Utterances halting, fragmentary and incoherent
1	Utterances hesitant and often incomplete except in a few stock remarks and responses. Sentences are, for the most part, disjointed and restricted in length
2	Signs of developing attempts at using cohesive devices, especially conjunctions. Utterances may still be hesitant, but are gaining in coherence, speed and length
3	Utterances, whilst occasionally hesitant, are characterised by an evenness and flow, hindered, very occasionally by groping, rephrasing and circumlocutions. Inter-sentential connectors are used effectively as fillers

Source: Weir (1993: 44)/TEEP, CALS, University of Reading

In the criteria shown, mention of actual language use is limited to *cohesive devices, especially conjunctions* and *inter-sentential connectors used effectively as fillers*; otherwise all references are to what are sometimes referred to as 'temporal variables' (Towell *et al.* 1996: 90) or as 'speech-pause relationships and frequency of occurrences of dysfluency markers' (Lennon 1990: 388). While these variables are of little value in terms of feedback to learners, they may provide useful benchmarks for raters. Some recent studies have set out to test the extent to which measures of these variables apparently do give true indications of fluency.

Lennon (1990) has attempted to establish reliable markers of fluency, investigating ten objectively measured variables, typically associated with fluency. His data is taken from a group of four students of EFL, who gave speaking samples before and at the end of a six-month interval of residence in Britain. Ten raters generally agreed that the second set of recordings was more fluent than the first. The variables measured in the two samples were all to do with either timing and the relationship of speech to pauses or with the frequency of markers of dysfluency, i.e. repetitions and self-corrections. On the basis of what Lennon found to differentiate between speakers at different stages in acquiring fluency, he concludes that, for the particular learner group, improved fluency was associated with:

- reduction of filled pauses and repetitions
- speech-rate improvement
- reduction of pause time (judged on the basis of improved speech-time ratio, increased length of runs between pauses (measured in words) and fewer disruptive internal pauses) (1990: 414).

Lennon further maintains that the improved rate of speaking was a function of reduced pausing rather than of faster articulation. He also concludes that self-correcting proved to be a poor indicator of dysfluency, and may even be regarded as a marker of fluency development.

Freed (1995) has carried out a study with elements resembling those cited from the studies of Lennon (1990) and Esser (1996). Freed's aim was to establish what it is in the speech of students who have studied French abroad that makes them sound more fluent than those who have studied at home. Six raters judged the fluency of oral language samples, taken at the beginning and end of a semester (pre- and post-test), from 30 students, half of whom spent the semester in France. The ratings suggested that, at least in the case of initially less advanced students, greater gains in fluency were made by those students who studied abroad than by the home students. The raters were asked, both in open-ended and closed-ended questions, to identify aspects of performance that influenced their judgement of fluency, and, as in the case of Esser's study, a range of variables was offered that included (but went far beyond) those temporal variables most commonly associated with the narrow concept of fluency.

Like Lennon, Freed has carried her research a stage further to investigate whether certain measurable attributes can be found that distinguish between groups of learners perceived to be at different levels of fluency. An analysis was carried out of the post-test language of eight of the students, who had been given similar, lower to mid level, fluency ratings in the pre-test, and of whom half had spent the semester abroad. Seven variables were measured and compared across the two groups (home and abroad). These variables covered: amount and rate of speech; frequency of unfilled pauses which were regarded as dysfluent (i.e. not occurring at clause boundaries); frequency of filled pauses; length of fluent speech runs (in words); repairs and clusters of dysfluencies. The only variable on which a significant difference was found between the groups was the greater rate of speech in the abroad group, but tendencies in the data suggested that other variables differentiate between the groups. The more fluent group seems to be characterised by:

- increased speech-rate
- more speech
- fewer dysfluent silent and non-lexical-filled pauses
- longer uninterrupted speech runs
- more repairs.

Despite this apparent evidence that certain measurable features act as indicators of fluency, Freed concludes by emphasising that raters, by their own accounts, do not judge fluency by these temporal qualities alone. She maintains that these observations 'bolster previous discussions describing fluency as a simultaneously vague and complex notion which includes a constellation of interactive features', and goes on 'Perhaps Sajavaara's recent observation that "fluency is ultimately in the ear of the listener" (1994: personal communication) is the most apt summary of our judges' evaluations' (1995: 143).

Towell *et al.* (1996) also address the question of which temporal variables act as indicators of fluency, considering similar variables to those cited in the studies above, but with a different motive for doing so. Rather than seeking to establish fluency markers as such, for surface recognition, they take a psycholinguistic approach, using their findings to shed light on how fluency comes about. Referring to both Anderson's (1983) ACT model of cognitive development and Levelt's (1989) model of speech production (to be studied more closely in 'Levelt's perspective: speech production and fluency' pages 148 – 151), they maintain that fluency in language production depends on linguistic knowledge being proceduralised, or automised.

According to the Levelt model, speech production takes place in three stages: in the conceptualiser (when we decide what we want to say), in the formulator (when we decide how to say it) and in the articulator (when we say it). Towell *et al.* hypothesise that an increase in fluency in advanced second-language learners is mainly attributable to a greater store of 'proceduralised knowledge' to draw on at the formulator stage. In other words, while the more fluent of these learners will not necessarily be more adept at thinking what to say, or at actually articulating, they will make significant savings in the time spent deciding how to formulate their ideas. This is because the proceduralisation of linguistic knowledge enables speakers to draw on ready-made units or 'productions', instead of having to construct all formulations from scratch, using 'declarative knowledge' about language. Procedures of the type *IF x, THEN y* (e.g. IF wishing to express doubt, THEN retrieve the unit 'I'm not sure', or just 'well ...') can be employed to carry out the formulation automatically. An increased use of these procedures in formulation frees the speaker to concentrate on other aspects of speaking, such as planning what to say, and so increases the overall rate of speaking. This speaking rate is assumed by the authors to be a global measure of fluency (c.f. the studies cited above), encompassing the whole working model of speech production (1996: 92). Towell *et al.* believe that speech drawing on a greater store of proceduralised knowledge in the formulator will be characterised by more and longer unbroken 'chunks'. They thus hypothesise that the greater overall speech rate of more fluent learners will be largely accounted for by an increased mean length of run between pauses.

They set out to demonstrate this by showing that, in the speech of learners whose fluency is perceived to have improved, an overall increase in speaking rate (SR) (syllabus/minute, including pauses) will be found to be accompanied by a significant increase in the mean length of run of unbroken speech (MLR) (in syllables). To eliminate the possibility that an increased MLR may be brought about by a greater time spent planning each run of speech, Towell *et al.* add the condition that the average length of pause (ALP) should not be seen to increase, nor should the phonation/time ratio (PTR) (actual speaking time/total time used to produce a speech sample) be found to decrease.

By comparing the speech samples from 12 advanced-level students of French after a period of residence abroad, with a similar sample taken before, Towell *et al.* found significant increases in the speaking rate (15 per cent) and the mean length of run (23 per cent), but no significant increase in the average length of pause, or decrease in the phonation/time ratio. The articulation rate (syllabus/second, excluding pauses) had increased somewhat (8 per cent), showing that some development had occurred at the articulator stage. However, the figures clearly indicate that the increased speaking rate is largely accounted for by the high increase in mean length of run. This then corroborated the hypothesis that, in the case of these learners, the increased fluency, as indicated by an increased overall rate of speaking, was mainly attributable to longer unbroken speech runs. This in turn suggests that a greater store of proceduralised linguistic knowledge is being drawn on in the formulation stage of speech production. Summarising Towell *et al.*'s findings along the lines of those cited earlier in this section, fluency can be said to be characterised by:

- speech-rate improvement
- longer uninterrupted speech runs.

The findings from these studies by Lennon, Freed and Towell *et al.* lead to some significant conclusions, insofar as they give empirical support to the notion that measurable indicators of fluency exist. While the only variable found in all three studies to differentiate significantly between more and less fluent speech was the overall rate of speaking, i.e. the amount of actual language produced in a given period of time, other variables recurred as apparent indicators of fluency. These include, notably, the mean length of unbroken speech runs (in words or syllables), usually combined with a decrease in the time spent pausing or in the frequency of pauses, particularly those regarded as disruptive (i.e. those that interrupt the natural flow of speech, by occurring at places other than the boundary of some kind of unit of information). While raters do not normally have access to the machinery required to measure these temporal variables exactly, they are all recognisable to the (trained) naked ear, and therefore have potential as factors to be included in scoring instruments, for the purpose of enhancing their reliability.

While the research cited here has revealed variables that can be useful in validating scales of speaking (the frequencies of the variables can be correlated with the grades awarded), they do little for the actual building of scales. This is partly because they do not identify the way individual features differentially discriminate between performances at specific levels of fluency; they rather establish that these features together discriminate between learners that 'have' (usually after a period abroad) or 'have not' got fluency. Moreover, they do not cater for the fact that pausing and the 'breaking up' of speech is, in fact, a very natural part of the speech of fluent and native speakers.

Fulcher (1996) maintains that empirical work on the concept of fluency is 'limited and inconclusive', adding 'the definitions of fluency which exist seem to be inadequate for the purposes of operationalisation in a test, even though the concept is widespread in the literature' (1996: 210). He goes on to illustrate the way fluency scales have traditionally focused on hesitation, making the seeming assumption that there is a 'motonic development of fluency from 0 to "perfect"' (1996: 210), characterised by a steady linear decrease in hesitation.

Fulcher (1993, 1996), questioning this assumption, makes a detailed study of the transcripts of 21 students across five oral proficiency levels, noting the frequency of pauses associated with the following eight contextual categories:

- end-of-turn pauses
- content-planning hesitation
- grammatical planning hesitation
- addition of examples
- expressing lexical uncertainty
- grammatical and lexical repair
- expressing propositional uncertainty
- misunderstanding or breakdown in communication.

Fulcher is able to identify the correspondence between the pausing behaviour within each category and a level on a five-band oral fluency scale, developed directly as a result of the research. This yields a detailed description of fluency at the various levels, which describes language behaviour in terms of pausing in very different ways, some of which are highly characteristic of the speech of native speakers. This makes a considerable contribution to the understanding of what goes into making speech fluent, and gives very concrete criteria for placing candidates on a scale of fluency. Salient characteristics of performance at five levels of fluency as described in Fulcher's terms (1996: 235 – 238) can be roughly summed up as follows:

band 1

The least fluent speakers are characterised by very short (often one-word) utterances, and lack of comprehension and of the ability to get clarification.

Long pauses occur when trying to understand the interlocutor or in searching for word or forms, and repetitions and restarts are common. The speaker is sometimes unable to make a response, and messages are sometimes abandoned because of language shortcomings.

band 2
At this level, speakers will still frequently need help in order to understand their interlocutor, but messages, once started, are generally fulfilled in a simplistic way without expansion, e.g. through examples. Pausing still occurs when looking for lexical or grammatical choice – sometimes with circumlocutions and often with midway switch of formulation.

band 3
Speakers usually understand the interlocutor. They seem more aware of the proposition, and spend time planning this. Appropriacy of word choice becomes more important, and pausing will occur in making these choices, with some appealing to the interlocutor. Utterances tend to be more expanded. Back channelling – using *hm* or *yeah* – helps to make conversation more natural.

band 4
Misunderstandings are rare. Speakers use hedges to express lack of certainty in the propositions. Few single-word utterances are given, and speakers expand their utterances, e.g. providing back-ups to opinions. Time is spent planning the content of the proposition and on how exactly to express themselves and present their views. Reformulations occur when the speaker is not satisfied with the proposition or with the correctness of the formulation.

band 5
Speakers demonstrate more confidence and are less likely to express propositional uncertainty. They rarely pause for reasons of grammar or word choice. Reformulations occur mainly for reasons of expressing the proposition fully. They expand and support themes. They respond very quickly.

Fulcher's characterisation of speech at different fluency levels is of particular value in that it contains reference to a variety of specific types of recognisable behaviour. Although references to actual words and expressions are restricted to occasional backchannels, or uncertainty markers, Fulcher paves the way for a 'fleshing out' of the language behaviour at different fluency levels with reference to the use of particular linguistic forms.

Further work towards the identification of 'linguistic' fluency markers is done by Towell *et al.* (1996), whose contention that the proceduralisation of linguistic knowledge is a significant source of fluency is cited above. Towell *et al.* maintain that this proceduralisation results in the use of ready-made units

of language. This revelation has potential in that it can be followed up by a search for evidence of the types of ready-made language units typically found in more fluent speech.

Towell *et al.* themselves suggest an approach that may be taken in this quest, involving Nattinger and DeCarrico's (1992) categories of syntactic strings, collocations and lexical phrases, which, they believe, pertain to issues that 'deal directly with the formulator rather than with other parts of the [Levelt] model' (1996: 105). They illustrate how these categories can be used in analysing speech samples, in order to investigate which types of linguistic knowledge are proceduralised by more fluent learners.

A more conclusive study of 'what is proceduralised' is that conducted by Raupach (1984) on the language of German students of French. In a case study of the language of two students, Raupach reports that uninterrupted speech segments are significantly longer in the language samples collected after a period in France. Raupach goes on to state that 'this is in full agreement with nearly all our second language data collected before and after a learner's stay abroad ...'(1984: 131). As has been shown above, this discovery is supported by Towell *et al.*'s (1996) findings.

Raupach ascribes this lengthening of unbroken speech runs largely to the acquisition of formulae, or recurring chunks. His study focuses on those formulae that tend to occur in combination with hesitation devices, suggesting that planning takes place while making use of these devices. He uses two categories of formulae. The first are 'fillers and modifiers', such as *je ne sais pas* and *je crois*, 'which do not have an immediate impact on the structure of the utterance "in process" but which, among other things, serve to give the speaker additional time for his planning activities' (1984: 123). The second are 'organisers', such as *c'est* and *on peux dire*, 'which contribute to the development of ongoing speech in that they help the speaker to structure his performance on the text level as well as on the sentence or phrase level' (1984: 123).

Raupach's analysis shows that the students, prior to their stay abroad, used many more one-syllable fillers, such as *euh* (non-verbal) and *et,* and generally displayed highly idiosyncratic planning behaviour, resulting in broken, dysfluent speech stretches. However, after their stay they showed near-native segmentation of speech stretches, due largely to idiomatic use of formulae as planning devices. By adopting formulae to 'fill' the place of a silent or unfilled pause, Raupach noticed the speakers used these in places more appropriate to the L2, helping them to abandon the 'temporal patterning' of the L1. Raupach goes on to comment, however, that the use of these is remarkably similar across learners, in contrast to their command of vocabulary, and he makes the point that the learners tend to restrict themselves to a narrow repertoire of organising formulae, using certain ones excessively frequently, sometimes to the extent that performance is perceived as non-idiomatic.

A language of fluency?

Raupach's analysis of language (carried out in the largely pre-corpus era) involves samples from only two students, and is fairly unrestrained in what is accepted as a formula (basing this largely on a fine segmentation of speech chunks by prosodic boundaries). However, it provides some empirical evidence that prefabricated chunks of language, and specifically those that are very frequent and colloquial and have little semantic meaning, co-occur with what have been shown, in this chapter, to be fairly reliable temporal markers of fluency.

The notion that the use of formulaic, or proceduralised, speech contributes to fluency is by no means novel. Pawley and Syder (1983) maintain that 'lexicalised sentence stems and other memorised strings form the main building blocks of fluent connected speech' (1983: 214). Bygate (1987) highlights the need for a stock of devices for facilitating speech, routines for structuring speech and procedures for negotiating meaning. Sinclair (1991) proposes the 'principle of idiom' by which 'a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices', adding that this 'may illustrate a natural tendency to economy of effort; or it may be motivated in part to the exigencies of real-time conversation' (1991: 110).

Furthermore, within this larger category of chunks, denoted by whatever name, an area of language roughly corresponding to the fillers, modifiers and organisers highlighted in Raupach's study has been explicitly associated with fluent speech. Nattinger and DeCarrico (1992) identify a sub-group of lexical phrases as 'discourse devices', with a further sub-category of 'fluency devices', such as *you know, I think,* and *at any rate* (1992: 64). Stenström (1994), using corpus linguistics, gives inventories of 'interactional signals' (e.g. *well, I mean, you know*), which 'play a crucial role in smooth interaction' (1994: 61) and 'discourse markers' (e.g. *right, well, anyway*), which 'help the speaker organise the discourse' (1994: 63).

What is emerging here is quite clearly a recognition in the literature (drawing on case studies, corpus linguistics, psycholinguistic models and theoretical reasoning) that fluency in speaking is enhanced by the use of ready-made chunks of language, and, more specifically, by the *smallwords* of speaking, as they have been defined in this study. This conclusion underscores what has already been noted in the introduction to this chapter, viz. that smallword use is instrumental in putting into effect all the skills identified as being specific to fluent speaking, principally by reference to Bygate (1987).

Fluency summarised

This section has taken the bull of fluency by the horns and attempted to describe what it is and how we might recognise it. A number of works have

been consulted, which reveal both the intangible and, happily, the more tangible sides of fluency. It has been shown that raters, while having some feel for what fluency is, tend to be vague and idiosyncratic when it comes to describing it. In attempts to give raters a handle on what this elusive thing is that they are supposed to be assessing, a number of the more core conceptions of fluency – *flow, unbrokenness,* and *smoothness* – have been translated into a set of variables, such as rate of speech and frequency of pausing. The reliability of these variables has been empirically investigated and a number of them shown to be fairly true indicators of fluency, notably the reduction of disruptive pauses, the rate of speech and the mean length of runs of unbroken speech.

The work of Fulcher (1993, 1996) adds to the understanding of the relationship between pausing and fluency, and identifies specific categories of language behaviour which characterise fluency in performance. These can be summed up as follows:

- pausing for content rather than language planning
- more complete and expanded messages
- expression of uncertainty of the proposition, rather than the language used
- repair and clarification to avoid potential misunderstanding
- quicker and more confident response.

As the present study is particularly concerned with finding linguistic markers of fluency, i.e. actual areas of language use associated with fluent speaking, studies were consulted that put more focus on the language forms used by speakers at more and less fluent stages. In line with the belief that fluency is enhanced largely by drawing on an increased store of proceduralised linguistic knowledge, the studies of Raupach (1984) and Towell *et al.* (1996) suggest that more fluent speech is characterised by more widespread, idiomatic use of ready-made chunks. Of these chunks, a group that can be equated with smallwords has been explicitly demonstrated to contribute to fluent speech. This association between smallwords and fluency derives considerable support from the literature on the process of speaking. As a final summing-up, fluency is defined for the purposes of this study as:

> the ability to contribute to what a listener, proficient in the language, would normally perceive as coherent speech, which can be understood without undue strain, and is carried out at a comfortable pace, not being disjointed or disrupted by excessive hesitation.

Principal recognisable markers of fluency in speech are:

temporal variables

- increased overall rate of speech
- increased mean length of unbroken run of speech

- decreased frequency of disruptive unfilled or non-verbal filled pauses

and

linguistic variables

- increased nativelike use of formulaic expressions generally
- increased nativelike use of smallwords.

The question of how smallwords function in bringing about fluency is the principal subject of the remainder of the chapter. Here, the notion of fluency will be recognised as being more than a set of tangible markers, and an attempt will be made to restore the concept of fluency as being an interaction between the listener and what s/he hears.

Forging a link between smallwords and fluency

The above definition of fluency presents the way fluency manifests itself in communication. What needs to be addressed now is the underlying means of achieving fluency, and more specifically the part played by smallwords. Smallwords have already been identified as linguistic markers of fluency, noted by many researchers as occurring increasingly as fluency advances. Whether or not this is corroborated by the data compiled in this study will be further investigated in the forthcoming chapters. However, before embarking on such an investigation, it is necessary to establish how smallwords work and why they should actually cause fluency (and not the other way round). Only after this is done will we be in a position to analyse the smallword use of learners and to draw meaningful conclusions about the effect of this on their fluency. Smallwords will therefore first be put under the spotlight, in an attempt to uncover the way they work in the cause of fluency, and whether there is a systematic and justifiable way of judging 'whole' smallword use among learners as part of the assessment of their fluency.

Smallwords in other people's books

Smallwords have been given a working definition in this study as:

> small words and phrases, occurring with high frequency in the spoken language, that help to keep our speech flowing, yet do not contribute essentially to the message itself.

In this section, some of the ways in which recent researchers have explained the workings of these small words and expressions will be looked at, with fluency in mind, as it was defined above:

> the ability to contribute to what a listener, proficient in the language, would normally perceive as coherent speech, which can be understood without undue strain, and is carried out at a comfortable pace, not being disjointed or disrupted by excessive hesitation.

Coherence, according to Crystal (1991), refers to 'the main principle of organisation postulated [in discourse analysis] to account for the underlying functional connectedness or identity of a piece of spoken or written language' (1991: 60). Stenström (1994) clearly assigns the role of organising and connecting spoken texts to smallwords, particularly those belonging to the two wide categories of what she refers to as 'interactional signals' and 'discourse markers'. Interactional signals (1994: 61) consist of those smallwords that do across-turn connecting work, involving or acknowledging the interlocutor, or what s/he says, such as *really?* and *you see,* while discourse markers, such as *anyway* and *right*, help the speaker to organise the discourse within his/her turn (1994: 63). At a narrower level, Stenström identifies numerous jobs done by smallwords. *You know,* for example, may perform the act of 'empathising' (which contributes to coherence at the level of the participants' shared beliefs), while 'frames', such as *all right*, and 'prefaces', such as *what else*, play roles in structuring conversation.

The role of smallwords in making speech understandable without undue strain is also reflected in Stenström's account. 'Hedging', e.g. *sort of*, may take the burden of literal interpretation off the listener, while 'monitors', such as *I mean*, help fend off possible misunderstandings.

It is clear from Stenström's account that the coherence brought about by smallwords is not restricted to connecting or organising text in its narrowest sense – i.e. what is said – but also in its broader aspects, such as whose turn it is, what ideas are being put across and which acts the speakers are performing through speaking.

Schiffrin (1987) sets out to show that coherence is brought about by 'discourse markers', defined as 'sequentially dependent elements which bracket units of talk' (1987: 31), such as *well, oh* and *I mean*. She maintains that coherence is created in the minds of the speakers and hearers through the integration of five planes of discourse:

- exchange structure
- action structure
- ideational structure
- participant framework
- information state (1987: 25)

In other words, what we say has simultaneously to fit into an exchange pattern, perform some kind of act(s), make sense within the ideas being developed, acknowledge the participation of others and take heed of the ever-changing information state of the participants. Schiffrin's basic contention is that

discourse markers play a crucial role in contributing to coherence by 'locating utterances on particular planes of talk' (1987: 326).

What the marker, or smallword, contributes to the discourse is maintained by Schiffrin to be a function of its inherent meaning and its location with respect to the various planes of discourse. For example, she concludes that the marker *oh* (inherently) 'marks a focus of speaker's attention'. Furthermore, if uttered on its own as a 'backchannel' (see 'The work of smallwords in optimalising fluency' pages 142 – 148), '*oh* ratifies the current participation structure of the conversation: speaker remains speaker', while, at the same time, it 'marks information receipt' (1987: 99).

Schiffrin also maintains that markers can be interchangeable on particular levels of speaking, but that they will not be equivalent on all other planes. This can be illustrated by considering *oh* and *right*. Stenström (1994) categorises both of these as 'acknowledges' when they occupy the 'follow-up' slot in the exchange structure. However, even in this role they are not equivalent on the information state plane, where *oh* suggests a change in information state, and where *right* (which Schiffrin does not actually analyse) is more suggestive of the fact that the information state has been confirmed, or adjusted in a way that does not dramatically add to or conflict with the earlier state.

Schiffrin's conception of coherence, besides involving several planes of discourse, may be interpreted as involving the comprehensibility of discourse. She maintains:

> Coherence then, would depend on a speaker's successful integration of different verbal and non-verbal devices, to situate a message in an interpretive frame, and a hearer's corresponding synthetic ability to respond to such cues as a totality in order to interpret that message. (1987: 22)

This view of coherence as implying successful interpretation of the message is compatible with the way fluency has been discussed and defined here. Thus support can be derived from Schiffrin for the claim that smallwords play a major part in contributing to fluency. Or, in Schiffrin's own words:

> Only one linguistic item – the discourse marker [...], anchors an utterance into more than one discourse component at once. By doing so, it provides a path towards the integration of those different components into one coherent discourse. Another way of saying this is that markers allow speakers to construct and integrate multiple planes and dimensions of an emergent reality; it is out of such processes that coherent discourse results. (1987: 330)

What both of these studies have shown is that smallwords are crucial to making spoken discourse cohere and make sense. Stenström illustrates how smallwords perform acts that contribute to the system of moves, turns, exchanges and transactions that make up the organism of spoken interaction.

Schiffrin takes a group of core smallwords – or discourse markers – as her starting point, showing how each of them performs, often simultaneously, the tasks of locating the speech on multiple planes of discourse. She illustrates how the various markers, while each bringing some inherent meaning to the discourse, take on a series of superimposed meanings from the plane(s) they are located on at the time. In doing so she highlights the efficient and systematic way smallwords work together.

Both authors, therefore, provide a convincing explanation of the way this small body of language plays an essential part in achieving fluency in speaking; moreover, they supply a mine of valuable information about how individual smallwords are used by native speakers, from both the USA and the UK. However, the very detail of their descriptions renders both works unsuitable as models for the simple, usable, unified framework sought here for analysing the roles played in bringing about fluency by any smallword in any speaking situation. To meet this demand, another work is consulted: Sperber and Wilson's (1995) presentation of 'relevance theory'. Here smallwords are barely mentioned, but the authors' account of verbal communication, I believe, provides the basis for a manageable way of describing the role of smallwords in bringing about fluency.

Smallwords and fluency in relevance-theory terms

Relevance theory, as posited by Sperber and Wilson (1995), is a theory of how people communicate verbally. It will be argued here that successful communication, as it is explained by relevance theory, corresponds closely with what has been defined here as fluent speaking. For one thing, relevance theory can be regarded as 'listener oriented', in that it focuses on how a listener interprets – through inference – what is being communicated, rather than on what is actually said. Furthermore, it gives an explanation of communication which takes account not only of how successful discourse is coherently connected but also of how it comes to be interpreted with minimum effort.

There are, moreover, many precedents for using relevance theory to explain the way smallwords, under various names, work, such as Jucker (1993), Aijmer (1996) and Andersen (1998). Jucker, in his study of *well*, states 'Relevance theory, I believe, is the only theory that can account for all the uses of *well* on the basis of a general theory of human communication based on cognitive principles' (1993: 438).

In this section, I will first give a simplified outline some of the 'essentials' of what constitutes relevance theory, highlighting a set of factors necessary for successful communication, or fluency. I will then go on to demonstrate how this yields a system of parameters which can offer an independent explanation of how smallwords work in the cause of fluency.

The essence of relevance theory

Verbal (understood here as 'spoken') communication, according to Sperber and Wilson, is brought about by a speaker intending to convey some 'meaning' to a hearer so that, in some way, their minds are brought closer in line, or their MUTUAL COGNITIVE ENVIRONMENTS (1995: 41) are strengthened. A person's cognitive environment consists of all the facts and assumptions that are MANIFEST (1995: 41) to her/him, i.e. that s/he is potentially, although not necessarily consciously, 'aware of' and capable of conceptualising. Meaning is conveyed principally through the hearer's drawing inferences, interpreting the utterance in the light of things in her/his cognitive environment; the actual utterance rarely carries the whole message explicitly.

If what is communicated causes some positive change (i.e. in line with the speaker's intentions) in the hearer's cognitive environment, it can be said to have POSITIVE COGNITIVE EFFECT (1995: 263 – 266) or to be RELEVANT. The cognitive effect, or 'changes in the individual's beliefs' (1995: 265), may be the adding of new assumptions, the strengthening of existing ones, or the elimination of formerly held assumptions incompatible with the new ones. The purpose of this affecting of the cognitive environment may be a genuine need to communicate information, or it may have a more 'social' motivation – to lay out common ground for further communication (1995: 62).

Cognitive effect is brought about by the hearer's making inferences on the basis of what is said and of whatever s/he can derive from a context that s/he selects to make sense of what was uttered. In order for this to happen, the utterance must be such that the hearer is not only able to infer the information, or message, that the speaker intended to communicate, but also the fact that the speaker intended to communicate it. The former condition involves the 'informative intention' of the speaker, while the latter involves the 'communicative intention' (1995: 29). The informative intention may be simplistically regarded as the intention to put across the message, i.e. what we are communicating, while the communicative intention involves making it clear what it is that we are doing in the communication. Any covert 'hidden agenda' we may have, which we do not intend to make obvious, is not regarded here as being communicated.

Verbal communication is thus regarded as OSTENSIVE-INFERENTIAL COMMUNICATION (1995: 63), 'ostensive' because the speaker's utterance, or STIMULUS, makes it plain what s/he is intending to do in the act of communication, and 'inferential' because the hearer is expected to infer the speaker's meaning from what is said in the context of other assumptions that are mutually manifest to speaker and hearer.

In order for communication to be successful (and although it carries no guarantee of success, it normally succeeds) the hearer must be able to draw the

correct inferences. This means that s/he must decide which of the potentially many meanings of an utterance is the right one, which, in turn implies knowing how to select a context in which it makes sense. The decision-making involved here is guided by the two PRINCIPLES OF RELEVANCE, which state (1995: 260):

1 Human cognition tends to be geared to the maximisation of relevance
2 Every act of ostensive communication communicates a presumption of its own optimal relevance.

According to the first principle, a hearer will naturally try to make sense of any utterance, i.e. to assign to it a context in which it will be relevant. By the second principle, the hearer can trust that what has been said has been done so in the most relevant way, so that her/his first, intuitive, interpretation of it will be the intended one. These principles are not maxims which communicators are expected to follow; they are descriptions of the way the mind works in the process of communication. The creation of relevance is a property of the mind, which may be regarded as contributing to the cost-effectiveness of verbal communication as it has evolved, just as, for example, certain biological properties contribute to the cost-effectiveness of muscle movement. The speaker knows that the hearer will respond according to these principles, and is able to use this fact when shaping her/his message.

The 'presumption of optimal relevance' is further defined in the following way (1995: 270):

a) the ostensive stimulus is relevant enough for it to be worth the addressee's effort to process it;
b) the ostensive stimulus is the most relevant one compatible with the communicator's abilities and preferences.

The reference here to two criteria – the effort of processing what is communicated and its degree of relevance, or cognitive effect – is reiterated throughout the discussion on relevance, which, it seems, hinges on a fine balance of the two. In order to achieve successful communication, a speaker will aim at putting the least possible processing load on the hearer, while, at the same time, 'keeping the channel open' by maintaining the highest level of cognitive effect possible under the circumstances. Put simply, the speaker will say as little as s/he needs to get the desired message across.

Just how a speaker achieves this balance, according to Sperber and Wilson (1995), is the subject of a long and complex discussion, which this book does not pretend to encapsulate. However, five factors can be identified, which, according to the principles of relevance theory, are fundamental to the success of the communication. Firstly, the hearer must be able to work out the communicative intention of the speaker, which, at its most basic, involves

knowing if s/he intends to communicate anything or to yield the turn, and, if s/he wishes to communicate, what s/he wants to 'do' through the communication. Secondly, since the hearer has to select a context to make the right interpretation of what is being said, s/he needs to be able to find this context for interpretation, e.g. from what has just been said, either within the speaker's own turn or earlier in the exchange, or from some external source. Thirdly, since this context may be partly derived from the immediately preceding utterance, it is relevant to know how the information or ideas expressed in that utterance were received, i.e. the cognitive effect of what was said. And, fourthly, the explicature (defined as 'the logical development of the propositional form' (1995: 182)) may not be entirely obvious, so that the hearer must be able to work out what is being 'said', not only by assigning referents to and disambiguating what is said, but also by enriching the proposition, 'in the presence of semantically incomplete or manifestly vague terms' (1995: 189). This enrichment of the explicature by the hearer may include interpreting the degree of commitment of the speaker to what is being said and how literally or vaguely the proposition is intended to be understood. Finally, so that it is mutually manifest that the intended communication has been brought about, the speaker, as well as the hearer, must be made aware of this fact, if necessary by a verbal signal. And, as has been pointed out, no guarantee of success is carried, so that if a hearer is unable to make a clear first choice in interpreting meaning, the onus is on her/him to have the situation clarified. Thus, it is important that both speaker and hearer are aware of the state of success of the communication.

Thus, it seems, that the success of verbal communication, i.e. the hearer's ability to make the correct inference as to what the speaker means to communicate, with the minimum of cognitive effort, depends on how readily the hearer is able to identify the following five parameters:

1 the communicative intention of the speaker
2 the context for interpretation of the utterance
3 the cognitive effect of the previous utterance
4 the degree of vagueness or commitment in an utterance
5 the state of success of the communication.

If the speaker communicates the message in such a way that the five factors above are clearly identifiable, the hearer will be able to interpret what is said with relative ease. Moreover, the context in which the utterance is interpreted involves, through these five parameters, the turn-taking system, the action (or what is being 'done'), the exchange (or conversation) in process, and the current state of ideas and information. Thus, the hearer, in setting these five parameters, is able to locate the utterance on the five planes of discourse cited in 'Smallwords in other peoples' books' pages 135 – 138 (although no

one-to-one correspondence is suggested), which means that the conditions for coherence, cited by Schiffrin (1987), are satisfied. What we have here, therefore, is clearly a recipe for fluent speaking, as this is characterised here, in the ear of the hearer, as coherent and understandable without undue strain. Relevance theory is thus interpreted in this study as offering a system of five parameters which, irrespective of the actual utterances used, are fundamental to fluent communication. It remains to be shown, for the purposes of this study, that the speaker is able to facilitate the hearer's tasks of finding the right 'setting' for each of these parameters by the use of smallwords.

Proposing a role for smallwords in relevance theory

There are numerous occasions when a hearer will need no help in setting the parameters above when making an inference (particularly in the case of 'default' situations, such as when the context for interpretation is the initial context, i.e. what has just been said). On some occasions, a gesture, look or grunt from the speaker, or even the physical environment, will provide a necessary clue. However, there are times when these non-verbal clues will not suffice. Sperber and Wilson state: 'A speaker who wants to achieve some particular effect should give whatever linguistic cues are needed to ensure that the interpretation consistent with the principle of relevance is the one she intended to convey' (1995: 249). While ostensive-inferential communication is regarded as primary, coded communication (i.e. using words) is acknowledged as a 'means of strengthening ostensive-inferential communication' (1995: 63). It seems reasonable to suppose that this coding may include a set of specific cues to help in the process of interpreting utterances through the parameter-setting referred to here. This, I believe, is where smallwords come in.

The compact nature of smallwords, their high frequency and the fact that they occur as a limited group, would suggest that their use puts a low processing load on the hearer – a fundamental property of relevance. The next section will illustrate how the use of the system of smallwords seems to play a key role in facilitating spoken communication.

The work of smallwords in optimalising fluency

In 'The essence of relevance theory' pages 139 – 142, it was implied that a fluent speaker is able to facilitate the hearer's tasks of finding the right 'setting' for each of the five parameters listed. In 'Proposing a role for smallwords in relevance theory', above, it was maintained that, at times, this involves the speaker's giving a linguistic cue as to how to set one or more parameters. Furthermore, it was suggested that smallwords act as prototypical linguistic cues of this type. The following discussion aims to illustrate how this is done. Actual examples of speech taken from test transcripts are not

given here but will be presented in the analysis of smallword use in Chapter 9. The discussion here largely uses Stenström's (1994) terminology for describing the acts and signalling functions performed by smallwords.

THE COMMUNICATIVE INTENTION OF THE SPEAKER

The communicative intention involves, most basically, the speaker's intended interactional behaviour. S/he needs to signal for example whether s/he intends to carry on speaking or to yield the turn to the other party. A 'filler' may be used, such as *well,* or *sort of* (Stenström 1994:76), to make it clear that s/he is not yielding the turn, despite a temporary break in flow, the choice of filler depending on which other signals need to be given at the same time. On the other hand, an 'appealer', such as *all right* or *you see* (both with rising intonation) (Stenström 1994: 80), can signal the speaker's intention that she wishes the current hearer to take the word. This point can perhaps best be appreciated by considering what might be the case if these signals were omitted – the information in the communication would not be affected, but the hearer might be unsure as to whether the speaker had finished or intended to go on.

Moreover, the communicative intention can be considered to involve the kind of message the speaker wants to communicate, i.e. what s/he intends to 'do' through the utterance. If a speaker intends to 'inform', s/he may begin with *you see* (Stenström 1994: 90). Or when the speaker is doing his/her best to explain something, this may be signalled with *I mean.* These intentions are often fully in line with what the hearer expects, or wants to hear. The speaker may, however, be about to communicate something which is contrary to what the hearer expects or would have preferred. This intention needs to be signalled and is frequently done so, in the case of a response, by using *well* (see R. Lakoff 1973, Svartvik 1980, Stenström 1984, Yule 1996).

THE CONTEXT FOR INTERPRETATION OF THE UTTERANCE

The notion of selecting a context containing facts or assumptions that make an utterance relevant in the hearer's mind is fundamental to relevance theory. Frequently, this is the 'initial context', i.e. that created by the immediately preceding utterance (either uttered by the previous speaker or the current speaker), frequently interacting with, for example, shared knowledge or the physical environment.

The speaker has to take into account how accessible the context for interpretation is to the hearer, and give verbal 'pointers' if necessary. If a temporary 'break' is made with the initial context, in order, for instance, to add information which is necessary in order to pick up the thread, but which may seem ostensibly 'beside the point', *well* is typically used (Stenström 1994: 115). Where the break with the initial context is more radical, and the focus is entirely on 'what is to come', this shift may be signalled with *by the way*

(Stenström 1994: 157) . If the salient context is that of an earlier part of the conversation, which has been temporarily left, the return to that context is marked, for instance, by *anyway* (Stenström 1994: 160).

THE COGNITIVE EFFECT OF THE PREVIOUS UTTERANCE

Besides pointing to where the hearer should turn to provide a context for interpreting an utterance, smallwords can indicate how an utterance relates to the previous utterance, particularly (but not necessarily) when this was given by another speaker. In other words, the cognitive effect of a previous utterance on the current speaker is signalled. Sperber and Wilson identify three main types of positive cognitive effect: the strengthening of existing assumptions, the adding of new ones and the replacing of existing, contradictory assumptions (1995: 114). Put simply, for an utterance to be relevant, it has either to make us more sure of what we thought before, or add to it in some way, or actually persuade us that what we thought before was wrong. If the utterance has no positive cognitive effect, either it tells us nothing new or we simply don't 'buy it'.

Smallwords can signal different kinds of positive effect as well as the lack of this effect. *Right* and *okay* tend to signal that previously held assumptions have been strengthened by what has been communicated, or that new but 'unsurprising' assumptions have been made. *Oh* tends to signal that new, more surprising assumptions have been made, possibly eliminating previously held ones. *Well,* on the other hand signals 'hesitation, or doubt, or scepticism and so on' (Stenström 1994: 113); in other words, it indicates some conflict between the existing assumptions and what has been communicated, which may need to be resolved before any cognitive effect is registered.

THE DEGREE OF VAGUENESS OR COMMITMENT INTENDED

So far the discussion has only considered smallwords used in a 'detached' way, i.e. not as part of a proposition, but rather as a kind of appendage to an utterance, or inserted as a filler, with little effect on the propositional content of the utterance.

However there is a case, I believe, for claiming that certain smallwords can assist in the interpretation of the actual proposition of an utterance. This seems to belong to what Sperber and Wilson (1995: 189) term the 'enrichment' of the semantic representation of the proposition, which they principally discuss with reference to the degree of commitment of the speaker to what is being said, and how literally or vaguely the proposition is intended to be understood. This can be signalled through the use of smallwords belonging to what has been termed 'vague language' in 'Speaking' pages 43 – 46. These smallwords may indicate either the degree of commitment to a proposition, e.g. by using an epistemic modal such as *I think,* or the inherent vagueness of the proposition itself, e.g. by using *around, or something* and *sort of.* Sperber and

Wilson state: 'In the model of ostensive-inferential communication we are trying to develop, impressions fall squarely within the domain of things that can be communicated, and their very vagueness can be described' (1995: 59; my underlining). When we are presenting an impression of something, whether because that is all we have in our minds, or because our linguistic resources allow no more than this, it is important that our hearer is aware of this, and does not make a mistaken literal interpretation. Smallwords, in describing the vagueness of a proposition or the speaker's attitude or degree of commitment to it, thus play an important role in assisting the interpretation of the explicatures of utterances. At the same time, they give the speaker a means of expressing imprecise or loosely held propositions.

THE STATE OF SUCCESS OF THE COMMUNICATION

Responsibility for successful communication does not rest with the speaker alone. It was stated at the start of 'The essence of relevance theory' pages 139 – 142, that verbal communication is achieved by a speaker intending to convey some 'meaning' to a hearer so that in some way their minds are brought closer in line, or their mutual cognitive environments are strengthened. This strengthening of mutual cognitive environments is only possible, by definition, if the parties are mutually aware of it. Although successful communication is probable, it is not guaranteed (Sperber and Wilson (1995: 17)), and the speaker cannot take success for granted without some signal from the hearer. Acknowledgement that the speaker's message is getting through can be communicated by the hearer, using non-verbal signal, such as nods and grunts, but it can also be done concisely with words, particularly through 'backchannels' such as *right* and *I see*. Backchannels are described in Stenström (1994) as making interaction possible 'without proper turn-taking, namely in cases where there is a (temporarily) dominant speaker and the other party's contribution is reduced to so-called 'backchannels' (realised by items like *m, yes, oh, I see, really*) as a sign of attention' (1994: 1).

The current speaker may also contribute, through smallwords, to this safeguarding of successful communication. This may be done by appealing either for confirmation that his/her message is getting across, e.g. through *right,* with rising intonation, or for help in getting this message across, e.g. through *you know*.

The discussion here has shown that smallwords send explicit signals that guide the hearer onto the correct parameter settings and hence to the interpretation of an utterance, and so reduce her/his processing load. Moreover, the fact that the same smallwords have sometimes been used during this discussion to exemplify various different signals illustrates another important point, viz., that smallwords are multifunctional. A single smallword, e.g. *well,* may simultaneously be signalling the intention to take the turn, to

prepare the hearer for an unexpected response and to indicate scepticism towards what the last speaker has just said.

The very efficient signalling work of smallwords in thus facilitating verbal communication and so optimalising fluency can be summed up as a set of five tasks:

1 they express the communicative intention of the speaker, with respect to what is to be communicated and how it affects the interactional roles of the participants
2 they point to the textual context in which an utterance has relevance
3 they indicate the cognitive effect of the preceding utterance
4 they indicate the degree of vagueness or commitment
5 they indicate the state of success of the communication, acknowledging it or appealing for confirmation or assistance in bringing it about.

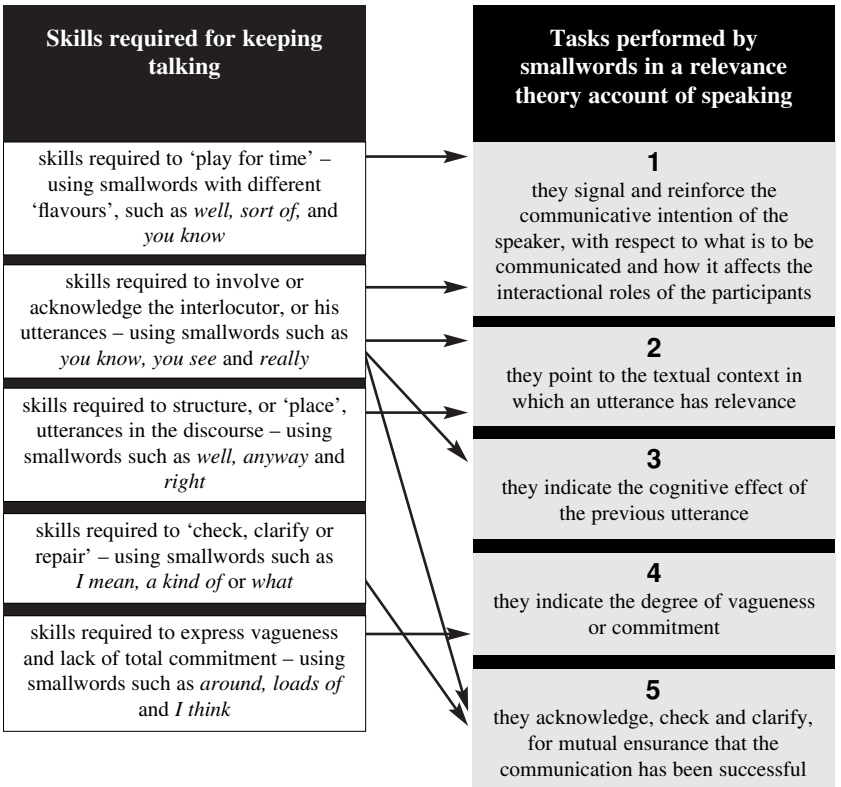
The five-task model outlined here is based on the way smallwords facilitate the processes involved in ostensive-inferential communication, as accounted for by Sperber and Wilson (1995). However, a close correspondence can be seen to exist between these tasks and the categorisation of the skills specific to speaking (based primarily on Bygate 1987), identified in 'Speaking' pages 43 – 46, which are associated with the preliminary definition of fluency given in 'The structural aspect of validity' pages 74 – 82. These skills were listed as follows:

- skills required to 'play for time' – using smallwords with different 'flavours', such as *well, sort of,* and *you know*
- skills required to involve or acknowledge the interlocutor, or his utterances – using smallwords such as *you know, right, you see* and *really*
- skills required to structure, or 'place', utterances in the discourse – using smallwords such as *well, anyway* and *right*
- skills required to 'check, clarify or repair' – using smallwords such as *you mean ...?, kind of* or *I mean*
- skills required to express vagueness and lack of total commitment – using smallwords such as *around, loads of, kind of* and *I think*.

A comparison of the two above listings reveals that the following correspondences, illustrated in Figure 6.1, can be made. Skills required to 'play for time', using fillers, correspond to task 1 in the 'smallword task list' – the signalling of the communicative intention – specifically to 'block' the possible interpretation that the speaker intends to yield the turn. The choice of filler will depend on which other signals the smallword is sending.

Skills required to involve or acknowledge the interlocutor, or his/her utterances, correspond to task 1 (when the communicative intention involves the signalling of the interactional roles), task 2 (where the salient context

Figure 6.1 The correspondence between skills specific to speaking (based primarily on Bygate 1987) and the five-part model of tasks performed by smallwords in a relevance theory account of fluency



may be part of the interlocutor's speech), task 3 (where the effect of the other speaker's utterances is indicated), and task 5, through backchannels.

Skills required to structure or 'place' utterances in the discourse can be regarded as corresponding to task 2, where the choice of smallword depends on whether or not an utterance can be interpreted directly in the context of the preceding one.

Skills required to 'check, clarify or repair' are associated largely with task 5, while skills required to express vagueness and commitment can be considered to correspond completely to task 4.

This close correspondence would appear to indicate that the above five-part model of tasks performed by smallwords, based on the relevance theory account of verbal communication, covers the skills identified earlier in

the study as associated with fluency, with no superfluous categories. This reinforces the belief that it is a model that accounts for the centrality of the work done by smallwords in bringing about fluency, as it has been described in two independent ways in this book. Having thus established that the model is a valid one for defining the work done by smallwords, it remains to build it up into a framework for analysing just how this work is done, i.e. which signals are being sent to indicate the actual settings on the five parameters in the model. Before proceeding to that step, however, I will consider another work which, I believe, reinforces the proposals made so far, by viewing communication from another perspective that of speech production.

Levelt's perspective: speech production and fluency

The discussion in the previous section hinged, deliberately, on how smallwords work to help a hearer to make sense of utterances. In this section, it will be demonstrated that the tasks performed by smallwords in affecting fluency are of as much service to the speaker as to the hearer. It will, moreover, raise the question of the extent to which foreign learners' fluency is dependent on smallword use. The discussion will centre on Levelt's (1989) model of speech production: *a blueprint for the speaker* (1989: 9) supplemented by De Bot's (1992) adaptation for bilingual speakers of Levelt's model.

The first question to be addressed here concerns why a speaker should wish to facilitate the hearer's drawing inferences from what s/he says and why we, as speakers, use inferential communication to the extent that we do – we could, after all, try to 'spell everything out'. Levelt (1989: 124) proposes four reasons for our using inferential communication. Firstly it is efficient – we save time and effort. Secondly it is a way of acknowledging the hearer's intelligence and cooperativeness, which leads to the third reason, that we wish to seem polite. The fourth reason is that to spell out what we mean in unnecessary detail would violate a maxim of quantity, raising questions in the hearer as to our motives, and leading to possible misinterpretation of our communicative intention.

In line with Sperber and Wilson (1995), Levelt maintains that 'the speaker's utterance invites the addressee to infer the communicative intention', and continues: 'Normally, the speaker's purpose will be that the hearer's representation agree in essential points with her own' (1989: 114). In other words, it is in the speaker's interest to communicate her/his meaning, and to do so ostensibly and through inference. Having established this, it follows that the speaker will be interested in ensuring that the hearer has the necessary clues to interpret the inferences, and that s/he does not expend unnecessary effort in the process – a struggling addressee can be as effort-consuming for the speaker as for the hearer. The speaker is not acting out of

benevolence, but has everything to gain by using ostensive-inferential communication, with all this implies, and by facilitating the hearer's interpretation of his/her utterances.

Levelt's model of the process of speaking is composed of a number of 'compartments' in which the different processes – from intention to articulation – take place. The processes are believed to take place incrementally during speaking, so that some parts of the utterance are being articulated while others are still at the various planning stages. For the purposes of the present discussion, two compartments will be looked at, these being the CONCEPTUALISER, where the pre-message is formed, and, to a lesser extent, the FORMULATOR, where this pre-verbal message is encoded. According to Levelt, the compartments cannot 'feed backwards', which entails that the conceptualiser is not open to 'protests' from the formulator, and will send out the preverbal message in its entirety to be put, somehow, into speech.

Two main levels of planning are identified within the conceptualiser: MACROPLANNING and MICROPLANNING. At the macroplanning stage, the communicative goals and series of sub-goals are determined, and the information to be expressed is retrieved. This can be regarded as the planning of the speech acts that are to be performed. Microplanning 'assigns the right propositional shape to each of these 'chunks' of information, as well as the informational perspective (the particular topic and focus) that will guide the addressee's allocation of attention' (1989: 11). So it is at the stage of conceptual microplanning – before any linguistic coding takes place – that the speaker decides which signals must be put into the utterance to lead to its interpretation. Levelt comments on the processes that go into microplanning as follows:

> A speaker will mark the referents in a message for their accessibility in such a way as to guide the listener's attention to what is already given in the discourse or to signal that a new entity is being introduced. He may also want to mark a particular referent as a topic. The speaker must further take care that all information is given the necessary propositional format, and that each pre-verbal message acknowledges the language-specific requirements of the Formulator. (1989: 108)

The main thrust of Levelt's message, as far as the present discussion is concerned, is that tasks of the kind identified in the previous section as being fundamental to the creation of optimal relevance, and 'performable' by smallwords, are laid down in the conceptualiser, in either the macroplanning or the microplanning stages, and put into the pre-verbal message.

On reaching the formulator, the entire preverbal message has to be encoded into 'words and grammar' and, ultimately, sounds. As was explained in 'Forging a link between smallwords and fluency' pages 135 – 151, citing Towell *et al.* (1996), two kinds of knowledge are drawn on at this stage –

DECLARATIVE KNOWLEDGE and PROCEDURAL KNOWLEDGE. Whereas the former is a body of knowledge we use to build up utterances, requiring some attention, the latter consists of fully automatic procedures, such as grammatical encoding or chunks of language that can be accessed 'ready-made'. Towell *et al.* maintain that, in the case of fluent speakers, this proceduralised, or automised, language knowledge includes formulae, and specifically smallwords, to carry out tasks such as organising or connecting discourse. Less fluent speakers, on the other hand, were found not to have access to this proceduralised body of knowledge.

The situation reported by Towell *et al.* seems explainable by Levelt's account, assuming that it applies to second-language learners as well as to native speakers. If the learner microplans in the way Levelt suggests, the pre-verbal message will be as complex, in terms of providing 'interpreting cues', as is the native speaker's. By the time a learner opens his mouth, he will have already given himself the need for *well, sort of, anyway, you know* or other expressions that make up smallwords. The conceptualiser will go ahead regardless, and if the learner has not proceduralised this body of knowledge, the formulation will be held up, interrupting the speech flow.

De Bot (1992) questions whether a literal application of Levelt's model, where compartments deliver to each other by a one-way system, can be applied to second-language learners. He maintains that, in their case, some form of 'warning' must be available at the conceptualiser stage, either from encyclopaedic knowledge (which includes his/her own access to lemmas) or by feedback from the formulator to the conceptualiser. This, he maintains, is necessary to prevent learners planning to say things they do not actually have the language to say.

If this should be the case, it would undoubtedly help the learner where, for example, a lexical or grammatical item is lacking, giving him the chance to adjust the pre-verbal message in microplanning so as to avoid that item. However, if smallwords are not accessible in the learner's proceduralised knowledge store, it is difficult to envisage which remedial alternatives are available.

This reference to Levelt's model has, I hope, succeeded in making several significant points. The first is that successful ostensive-inferential communication, based on the creation of optimal relevance in the hearer, is as necessary to the speaker as to the hearer. From this follows the next point that, in playing a major role in bringing about this success, smallwords are acting in the speaker's interests. It has also been shown that the clues that speakers give hearers for creating optimal relevance are determined in the pre-verbal message; according to Levelt, there is no way for the formulator to prevent them from being formed. If this is applied to second-language learners, they are clearly in continual need of smallwords, whether or not they have them in their store of procedural knowledge. And even if De Bot's (1992) idea of an

adjusted model for second-language speakers is accepted, it is difficult to see how a learner can adapt his microplanning for the lack of the very devices that are deemed to facilitate the complex ideas s/he needs to express.

Levelt's account of speech production can thus be regarded as reinforcing the account, developed in 'Smallwords and fluency in relevance-theory terms' pages 138 – 148, of the way smallwords facilitate the process of ostensive-inferential communication and hence contribute to fluency.

A framework for analysing smallword signals

In 'The work of smallwords in optimalising fluency' pages 142 – 148, a five-point model of the tasks performable by smallwords was worked out and presented in Figure 6.1 on page 147. Here the main signals, or 'macrosignals', of smallwords were established, such as pointing to the context for interpreting an utterance. Each of these macrosignals can be sent through particular 'microsignals' (often simply referred to here as 'signals') corresponding to the different parameter settings referred to in 'The essence of relevance theory' pages 139 – 142, e.g. signalling the intention to take the turn. In order to analyse the way individual smallwords operate, we need to establish, for each macrosignal, which are the principal microsignals potentially sent by smallwords. In this section, the five-point model is refined so that a framework of macro- and microsignals is produced for analysing the functions performed by smallwords in contributing to fluency.

The microsignals of smallwords are identified for each macrosignal in turn. These macrosignals are defined as follows, based on the five-point model presented on pages 142 – 148:

- expressing communicative intention
- pointing to the context for interpretation
- indicating cognitive effect
- indicating the degree of vagueness or commitment
- indicating the state of success of communication.

Each of these macrosignals is broken down into a series of microsignals to be identified in the data being analysed. Certain of the macrosignals may be considered to have 'default' settings, e.g. where the communicative intention of the speaker is that s/he is going to comply with what the previous speaker required (e.g. by giving a direct answer to a question). Default, or 'expected', signals have been identified here on the basis of which of the speaker's intentions can be considered, in the light of the discussion in this chapter, to be recognisable by the hearer with the least processing effort. In such cases it is reasonable to suppose that there is a reduced need for explicit signalling by a smallword, and that the onus on the speaker will rather be to signal an intention other than the default. For this reason, the microsignals identified

here weigh somewhat in the direction of non-default settings, which the hearer needs help in identifying.

The microsignals cited have been selected because, firstly, they seem to cover most of the prototypical signals associated with the five principal tasks of smallwords as they have been discussed in this chapter, and, secondly, because each is definable in a sufficiently tangible and distinct way as to allow a (relatively) uncontentious identification of evidence of its existence in the dataset of student test language to be analysed. The resulting taxonomy is presented at the end of this chapter and is discussed below.

EXPRESSING COMMUNICATIVE INTENTION

Two microsignals are identified within the macrosignal of expressing communicative intention. The first concerns whether the speaker intends to take, hold or yield the turn; as there is no 'default setting' of this particular intention, smallwords can be expected to be used in signalling any of these intentions.

The second concerns what the speaker intends to 'do' in the communication. The default setting is that the speaker intends to do what the hearer expects, i.e. by giving an anticipated or hoped for response. The non-default setting is where the speaker is unable or unwilling to give this hoped for or expected response. This kind of response is referred to here as an 'oblique' response.

The two microsignals associated with expressing communicative intention can be summed up as:

- signalling whether the speaker intends to take, hold or yield the turn
- signalling an oblique response.

POINTING TO THE CONTEXT FOR INTERPRETATION

The default context in which an utterance can be interpreted is the initial context (see 'The work of smallwords in optimalising fluency' pages 142 – 148), i.e. what has just been said, either by the same or by another speaker. Two scenarios of this not being the case are identified here as being signalled. The first is where the 'mode' of the speaking, used in the preceding utterances, is broken with in some way, so that the upcoming response cannot be interpreted directly in the light of that mode. (Here 'mode' is used to include the Hallidayan primary classification of discourse into speech and writing, but also sub-divisions of these, following Crystal 1991: 220.) This is most normally illustrated in the dataset being studied when the student takes the turn, switching from the 'real-life' situation of talking about the task to be done, and into the 'test' situation, embarking on the task, such as a narrative or description. This will be characterised here as 'mode changing'. The second is where the speaker breaks, in mid-utterance, with her/his own speech so that what s/he has just said no longer provides a relevant context for interpreting

what is to come. This may be in connection with a digression or a sudden disturbance to the flow of thought. Breaks due to self-correction or restarts are also included in this category.

The two microsignals used in pointing to the context for interpretation can be summed up as follows:

- signalling a break with the mode used in the preceding utterances (mode changing)
- signalling a mid-utterance break with context created by the speaker's own immediately preceding speech, including self-repair.

INDICATING COGNITIVE EFFECT

If an utterance is to have positive cognitive effect (i.e. be relevant) it must produce some change in the hearer's cognitive environment – i.e. the hearer makes some new inference from what s/he has heard. The default setting for the cognitive effect of an utterance is that this inference either strengthens previous assumptions the hearer has held or adds 'anticipated' assumptions (see 'The work of smallwords in optimalising fluency' pages 142 – 148). In other words, no particular surprise or emotion is evoked. This may be signalled by *right,* or *okay.*

When new 'unanticipated' assumptions are made by the hearer, or a conflict arises with the hearer's previous assumptions and the new inference is accepted, replacing the old assumption, this is frequently signalled by *oh.* This will be referred to as a 'change-of-state' signal, following Heritage (1984).

As an alternative to positive cognitive effect taking place, the new inference may be rejected by the hearer. As no instances were found in the data of smallwords signalling the rejection of new inferences, this microsignal will not be expanded on here, although it is included in the framework.

Thus, two microsignals are recognised here as being associated with indicating the cognitive effect of an utterance:

- signalling a cognitive 'change of state'
- signalling the rejection of the inferences of the previous utterance.

INDICATING THE DEGREE OF VAGUENESS OR COMMITMENT

While it might have been natural, on the basis of what was discussed on pages 142 – 148, to create two microsignals here – the signalling of vagueness 'within' the proposition, and the speaker's lack of commitment to the proposition – this has not been done. Instead, a single microsignal is identified within the category of indicating the degree of vagueness of an utterance or commitment to it, referred to here as softening the impact of a message, or 'hedging'.

This decision is partly influenced by Nikula (1996), who took a similar

approach to her study of 'pragmatic force modifiers' – as used by Finnish learners of English and native speakers. Nikula argues that it is unnecessary to make a distinction between epistemic modals, such as *I think*, and other kinds of 'vague language', such as *sort of*, since both have the effect of softening the impact of a message (1996: 46). She furthermore extends her concept of a pragmatic force modifier to include expressions such as *just*, which reflect the speaker's attitude to a proposition, either 'down-toning' or 'emphasising' it.

The term 'hedging' has been adopted here to describe the function of softening the force of a message, because it is already in widespread use, and because there are precedents in the literature for using the term to cover the types of softener referred to here. Lakoff (1982) defines the term 'hedge' as referring to 'words whose meaning implicitly involves fuzziness – words whose job is to make things fuzzy or less fuzzy' (1982: 195). While his list of examples and discussion of hedges do not cover epistemic modals, those of others do. Stenström (1994) says: 'by hedging, the speaker modifies what s/he says' (1994: 128) and includes epistemic modals in her examples, as do Brown and Levinson (1987: 162). Thus a single, inclusive, microsignal is recognised here as being associated with the vagueness or commitment in an utterance:

- signalling a softening of the message: hedging.

INDICATING THE STATE OF SUCCESS OF COMMUNICATION

Communication runs smoothly and successfully, on the whole, and the default microsignal within the macrosignal of indicating the successfulness of communication can be considered to be the acknowledgement that it is successful. This is done minimally through the use of backchannels, which 'signal listener attention' (Stenström 1994: 221). It is also carried out by the explicit signalling, e.g. in response to an appeal, that communication is running smoothly, without actually taking the turn or 'breaking the flow' of what the other speaker is saying.

At other times, the speaker finds it necessary to appeal to the hearer either to confirm that s/he is 'getting the point', or to help out in bringing about mutual understanding when the speaker has difficulty in expressing her/himself. This explicit 'appealing' (Stenström 1994: 221) must also be signalled. Thus there are two microsignals in this category:

- signalling the acknowledgement of smooth communication
- signalling an appeal to the listener to confirm or assist smooth communication.

Just how these different signals can be recognised in the dataset will be discussed in depth in Chapter 9 pages 224 – 240. Meanwhile, the framework to be used in analysing the macro- and microsignals that smallwords send in contributing to fluency is illustrated in Figure 6.2.

Figure 6.2 Framework for analysing smallwords in terms of the macro- and microsignals they send

MACROSIGNALS	MICROSIGNALS
SIGNALLING THE COMMUNICATIVE INTENTION	signalling whether the speaker intends to take, hold or yield the turn signalling an 'oblique' response
POINTING TO THE CONTEXT FOR INTERPRETATION	signalling a break with initial context created by previous speaker ('mode changing') signalling a mid-utterance break with context created by the speaker's own immediately preceding speech (including self-repair).
INDICATING THE COGNITIVE EFFECT OF PREVIOUS UTTERANCE	signalling a cognitive 'change-of-state' signalling rejection of the inferences of the previous utterances
INDICATING THE DEGREE OF VAGUENESS OR COMMITMENT	signalling a 'softening' of the message: hedging
INDICATING THE STATE OF SUCCESS OF COMMUNICATION	signalling the acknowledgement of successful communication signalling an appeal to the listener to confirm or assist successful communication

Summary

This chapter began with an attempt to pinpoint the nature of fluency. Based largely on the work of other researchers, a definition was gradually worked out, incorporating what appear to be generally acknowledged as core components of fluency. Moreover, certain markers were identified – some temporal, some linguistic – which seem to be reliably associated with the speech of learners at more advanced stages of fluency. Among the latter type,

smallwords emerged as making up a significant part of what might be regarded as the particular 'language of fluency'.

In order to provide a theoretical basis for claiming a causal link between smallwords and fluency, and to work out a theory-based framework for the analysis of smallword use in the current investigation, it was necessary to probe beneath the surface of fluency to see how it may be brought about and what role smallwords have in the process.

The works of Schiffrin (1987) and Stenström (1994) were consulted to obtain a deeper understanding of the work done collectively by smallwords, under various names, in promoting coherent and intelligible speech, and hence *fluency*, as it is defined here. However, it was Sperber & Wilson's (1995) *relevance theory* account of verbal communication that provided the foundation for the framework being sought after. This account was interpreted as revealing a set of five factors which contribute to the hearer's ability to readily make sense of what s/he hears. These five factors were all demonstrated to be capable of being brought to the hearer's attention by means of smallwords. Moreover, this five-factor model of communication was shown to explain not only ease of interpretation but coherence in Schiffrin's (1987) terms, and hence, by the definition arrived at in this book, fluency. It was also found to coincide by and large with the skills associated with fluent speaking, identified at a preliminary stage in this book and based largely on Bygate's (1987) account of speaking. A further perspective was lent to the study by consulting Levelt's (1989) account of speech production, as well as De Bot's (1992) application of Levelt's theory to non-native speakers. It was established that smallwords, in the roles assigned to them here, facilitate not only the interpretation of speech but also its production. The fluency-dilemma of the learner who does not have a ready store of smallwords to draw on was highlighted.

Finally, each of the principal tasks – in terms of the macrosignals that can be sent by smallwords – in the five-part model was further broken down into a series of specific microsignals. The resulting framework, built on a relevance theory account of fluency and reinforced by the recent findings of other scholars, will hopefully be proven to be a comprehensive, user-friendly framework through which all smallwords found in student data can be classified for the signal(s) they send in promoting fluency.

7 Smallwords and other fluency markers: quantitative analysis

In the previous chapter, we paved the way for analysing the signals sent by smallwords in the speech of learners and native speakers. However, before proceeding to find out how the students taking the EVA test used smallwords, it is necessary to establish the extent to which they used them. Only if it should transpire that the more fluent group of students were greater users of smallwords, as is hypothesised here, will the study of the manner of this use have relevance for this book. The question of the 'quality' of smallword use will therefore be left for Chapter 8 pages 183 – 223. This chapter addresses the more quantitative issues surrounding smallword use as well as other, temporal, indicators of fluency.

The overall aim in this chapter is to find out, through the analysis of transcripts, information about which features actually appear to distinguish degrees of fluency in the students' speech. This information is needed so that statements can be written, with some confidence, into scoring instruments about fluent performance at different levels.

In 'Fluency summarised' pages 133 – 135, certain recognisable variables were identified as markers of fluency in speech:

temporal variables
- increased overall rate of speech
- increased mean length of unbroken run of speech
- decreased frequency of disruptive unfilled or non-verbal pauses

linguistic variables
- increased nativelike use of formulaic expressions generally
- increased nativelike use of smallwords.

The temporal variables seem, on the face of it, unproblematic insofar as they are inherent to fluency as it has been defined here, and are associated with recurrent core elements in most definitions and descriptions of fluency. They are widely recognised as 'symptoms' of fluency, and, as long as they can be identified, they can be used to distinguish fluency levels between learners, as has been empirically shown in works cited in 'Pinning down fluency' pages 124 –125 (e.g. Lennon 1990). They would therefore appear to be reliable elements for inclusion in descriptors of fluency in performance. However, an

unresolved problem exists in the case of pausing. As Fulcher's (1996) findings (see 'Identifying elements of fluency' pages 126 – 132) have illustrated, pausing does not simply decrease linearly as fluency advances. The use of the term 'disruptive pause' here should go some way to limiting the concept of pausing as a temporal marker of fluency to the pausing associated with language planning or correction, typical of less fluent speakers. This pausing can be assumed to be disruptive since, unlike content planning, it will not normally occur at the boundaries of natural 'information units' (Stenström 1994: 8), and so will break up the message in an irregular way. However, this still leaves the problem that there is no manageable way of recognising all the disruptive pauses in the corpora. For the time being, therefore, the simplistic stance is taken that, under similar conditions, less fluent speakers will pause relatively more frequently than more fluent ones. Hopefully, an analysis of the way pausing is distributed across turn positions should shed light on the kind of pausing that appears to differentiate between the student groups.

The linguistic variables pose their own brand of problem, largely because they are not inherent to fluency as it is generally perceived. They do not noticeably feature in its definitions or scale descriptors, although, as has been demonstrated, they are increasingly acknowledged as contributors to fluent speech, and some empirical findings (see Raupach 1984) indicate that they do differentiate between more and less fluent speakers. Moreover, in the case of smallwords, a case was created in 'Forging a link between smallwords and fluency' pages 135 – 151, for arguing that the association between smallwords and fluency is a causal one; more nativelike use of smallwords brings about greater fluency.

Still, the case remains a tenuous one as long as more empirical evidence is not available. Doubt has been cast on the simple equation *more smallwords = more fluency* in Chapter 6, where the complexity of the ways smallwords function has been illustrated. Therefore, although it seems likely that more fluent learners use smallwords in a more nativelike way than do less fluent learners, a considerable amount of evidence is needed about not only the quantities of smallwords used, but also their range and distribution across turns in the speech of learner and native-speaker groups, before confident claims can be made or meaningful statements written into descriptors of performance.

This chapter, therefore, sets out to gather as much numeric information as possible about the use of smallwords and temporal markers by students of different fluency levels. The information on temporal markers should verify whether the students characterised as 'more fluent' (based on raters' judgements) really were more fluent, measured on core features of fluency. The information on smallwords should tell not only whether the stronger group used more of these, but also whether they used them to get started or to keep going, and whether they differed in their smallword choices. And both

types of information combined should reveal any relationship between pausing and smallword use. All of these issues are addressed in this chapter, and any features of behaviour that appear to differentiate students according to fluency are noted. As was explained in Chapter 6, the use of formulaic expressions is not investigated here, although it is to be hoped that findings on this will be available from other studies at a future stage to supplement the findings here. Nor are the categories of language behaviour identified by Fulcher (1996) as characterising speech at different fluency levels addressed at this stage.

The approach

The approach to this investigation is quantitative in that conclusions are drawn on the basis of comparing quantities of various potential fluency markers in the transcripts of groups of speakers. Three groups of speakers are involved: a more and a less fluent Norwegian group and a native speaker group. Certain countable features are looked for, such as smallwords and filled pauses, and their frequencies noted overall and in different turn positions. In the case of other features, such as the length of turns, group means are worked out and compared. The objectives are to test certain hypotheses and to explore the data for findings that may lead to a fuller picture of learner performance at different fluency levels.

What most strikingly differentiates this research from the empirical studies cited in 'Fluency' pages 124 – 135 (notably Raupach 1984; Lennon 1990; Freed 1995 and Towell *et al.* 1996) is perhaps the fact that corpus linguistics is used in this study; i.e. data from relatively large groups of students are accessed electronically here, whereas in the studies cited, language samples were generally obtained from a handful of students and studied manually. The advantages of using corpus linguistics are manifold. For one thing, speakers can be highly idiosyncratic in terms of both temporal variables and the use of smallwords. In restricted case studies, these idiosyncrasies can distort results to an extent that is unlikely when a dataset is made up of the transcripts of many speakers. Moreover, the greater quantity of occurrences of features means that more stable patterns can be observed, and that statistical tests can be used to support conclusions.

There are of course restrictions imposed by using corpora. These are primarily due to the fact that the information available is largely limited by what the corpus is 'tagged' for, i.e. the range of verbal or non-verbal features and variables, such as gender, that are coded for counting purposes. This tagging is done when the corpus is originally compiled and cannot normally be altered to suit the individual researcher's needs. Any additional information must, therefore, be found manually, e.g. by combing through

printouts or listening to tapes, and clearly, when many transcripts are involved, there is a limit to how much of this manual work can realistically be carried out. In the case of the present study, a restriction has been imposed by the fact that unfilled (silent) pauses are not considered to be reliably documented in the transcripts, so that findings on pausing are confined to those on non-verbal filled pausing. The discussion on the legitimacy of this is taken up in 'Findings on temporal variables' pages 165 – 170.

However, the limitations have not been so severe as to preclude essential components of the research, although these have sometimes had to be carried out in a roundabout way. Granger (1997) states 'SLA researchers should never hesitate to adopt a manual approach in lieu of, or to complement, a computer-based approach' (1997: 16–17). In this study, the computer-based approach would not have been possible without many hours of manual work, for instance, in deciding when a word that looks like a smallword actually functions as a smallword!

The method used is outlined below, but first an account is given of the data analysed and the smallwords searched for, as well as of the hypotheses and research questions behind the investigation.

The data

The data used in the quantitative analyses (and also in the qualitative analyses to be reported on in Chapter 8) consists of transcripts (and audio tapes) of students – 62 Norwegian and 26 English – carrying out the EVA speaking test, in pairs, with a tester. The transcripts form part of the EVA Corpus, which is made up of the recordings of all the EVA oral tests carried out during the national piloting, together with recordings of 'mock tests' made in two schools in the north-east of England. The Norwegian students who took the EVA test were selected mainly alphabetically from schools whose catchment areas represented a cross section of Norwegian society. The British students were selected by their teachers at schools with student intakes from very mixed socio-economic areas in the north-east of England. The teachers were asked to provide a representative mix of students with respect to ability and social class. All students were in the same age range, 14 to 15 years.

The transcripts from three groups of students have been selected for analysis: a 'more fluent' (NoA) and a 'less fluent' (NoB) Norwegian group (selected solely on the basis of global grades on the speaking test) and a native speaker group (NS) (selected primarily with gender balance in mind). The grouping has been designed to achieve a certain degree of parity between the groups, striking a balance between the numbers of students and the total numbers of words uttered. Additionally, as it is hoped that the findings of the investigations will lead to enhancement of the EVA speaking test assessment

scale bands, cut-off points for the grades defining the Norwegian groups are deliberately aligned with critical points on the band scales (see Appendix C pages 277 – 279).

The NoA group (19 students) is defined as all students with mean (across raters) global grades higher than 5.0 (i.e. from '5 plus'), in other words, true 'top band' students, while the NoB group (24 students) is defined as all students with mean global grades of less than 4.0 (i.e. from 4 minus down), which means, in reality, 'middle to low band' students; virtually no student was true lower band, as described on the scale, although there were occasional instances of students being awarded a 2 plus (14–15-year-old students in Norway rarely fit the descriptions on the lower bands of the speaking test scale, although they frequently do in the case of the writing test). Students ranging from grades 4 to 5 (19 students) were eliminated, to ensure that the ability levels of the two groups were truly distinct. One had to remember that the inter-rater reliability on overall grades left cause for concern and therefore raised questions of how 'true' grades were. Leaving 'clear water' of roughly one third of the students, around one and a half whole grades, between these groups was intended as a safeguard against rater effect. In the case of essential disagreement between the two raters (where a student was judged as strong by one and as weak by the other), students would normally land in the middle, excluded group. The raters would have had to disagree by about three whole grades for the average, i.e. final, grade to be so far out of line as to actually risk placing a student in the wrong fluency group as defined here, and in no instance did this happen.

It should be pointed out here that the global grade is not ostensibly a measure of fluency, but was the final grade arrived at by raters on the basis of setting a level on the two band scales (see 'Specifications for scoring procedures' pages 62 – 65). Unfortunately, no separate grades for fluency were submitted by raters. However, as was pointed out in 'Inter-rater reliability' pages 104 – 110, the four fluency-related sub-scores correlated highly (with r ranging from 0.59 to 0.78) with the mean global grade. What is more, raters were instructed to give their rating on the fluency scale priority over that on the language scale when setting the global grade. Thus it is considered justifiable to regard the global grade as a measure of fluency for the purposes of the analysis. However, in order to be completely satisfied that these groups are distinct in their levels of fluency, part of the process followed in the analysis is to find support for this grouping, by applying temporal measures of fluency.

The NoA group consists of 19 students, uttering a total of 14,066 words. The NoB group consists of 24 students, uttering a total of 10,467 words. In total, 22 boys and 21 girls are included (NoA 9:10, NoB 13:11). The NS (native speaker) group consists of 18 students uttering a total of 12,349 words, the students being chosen primarily with the gender balance in mind, so that

all eight boys in the dataset and the first ten girls on the list, sorted by ID coding, are included.

The transcripts have been computerised and tagged for the following variables: personal identity, gender, global grade (in the case of Norwegian students) and task (of which there are three in each test). Furthermore, the data has been processed to be accessed through a 'query form' using a TACTweb search programme, whereby words and phrases can be counted and 'called up' in context or normalised distribution, and analysed with respect to the tagged variables listed. No phonetic marking of any kind has been used, although unfilled pauses of different lengths have been marked. The voice recordings are not available 'on-line' (although they can be listened to separately) and there is, as yet, no means of measuring the rate of speaking. The transcription was carried out and proofread by bilingual English-Norwegian speakers (graduates). The electronic tagging was done professionally by the University of Bergen Humanistisk Datasenter.

In the later, qualitative analysis, in Chapter 8, possible idiosyncratic usage of smallwords will be checked for by comparison with other research findings on native speaker use of smallwords.

The smallwords

In Section 1.1, smallwords were given the following working definition:

> *small words and phrases, occurring with high frequency in the spoken language, that help to keep our speech flowing, yet do not contribute essentially to the message itself.*

While this definition, accompanied by examples, has been sufficient in the (so far) theoretical treatment of smallwords, there is clearly a need for a more precise definition if the use of these words and phrases is to be analysed in the transcript data. This has been attended to, not by inventing a new abstract definition of smallwords, but rather by defining them as a set consisting of certain members. This set is made up of all the words and phrases that occurred fairly commonly in the dataset, and which fitted the description above. The process of selecting the smallwords was as follows: firstly, a large section of the data (that of eight Norwegian and eight native-speaker students) was scanned manually for any word or phrase that was judged to qualify as a smallword according to the working definition. Those words and phrases were then isolated from the main dataset and studied further, in context, to assess their absolute eligibility as smallwords, and their frequency.

As most words and expressions are polysemous and have senses that do not conform to the smallword definition, occurrences with these senses had to be eliminated. The following occurrences of words and phrases were not counted as smallwords: *all right* as in *she's all right*, and uses of *like* as in *she looks*

like she's worried, and *I think*, as in *I think that it's her brother*. These words and expressions cannot be simply 'dropped' without fundamentally distorting the syntactic or semantic properties of the utterance. Occurrences of *just* that were clearly adverbial in function, e.g. *I'd just got there*, were also excluded. More detailed discussion of which occurrences 'count' as smallwords takes place in the relevant sections of Chapter 8.

In the case of smallwords with very few eligible occurrences, in the region of five or fewer, these were normally dropped from the study. However, in the case of a few of these low-frequency smallwords, which were formally and functionally so similar that the selection of one or other of them was probably idiosyncratic, these were put together in groups. Two such groups were formed on this basis: *and everything/and that/and stuff/and things* and *sort of/kind of*. Thus the following 19 smallwords (or smallwords groups) were yielded:

well	right	all right	okay	oh	ah
you know	I think	I mean	like	sort of/kind of	
a bit	just	or something	not really		
and everything/that/stuff/things			I know	you see	I see

The judgements on what counted as a token of a smallword were initially made by myself. However, a second native speaker was brought in, with occasional recourse to listening to tapes, in order to overcome some problems in deciding the eligibility of smallwords.

Hypotheses and research questions

Two main hypotheses are tested in this investigation:

- the frequency of temporal variables will be found to support the grouping, based on ratings, of the the Norwegian students into more and less fluent speakers
- the more fluent speakers will be found to have used smallwords in a more nativelike way than the less fluent, as far as quantity and distribution across turns are concerned.

The research questions address these hypotheses, in addition to attempting to find out anything that might indicate a link between smallword use and temporal variables. The following questions are posed:

1 Is there evidence to indicate that the NoA group was more fluent than the NoB group, judged by:
the occurrence of non-verbal filled pauses overall and in different turn-positions?
the mean length of turn?

2 Is there evidence to suggest that the NoA group used smallwords in a more nativelike way than the NoB group, regarding:
the total number of tokens, overall and in different turn positions?
the range (i.e. the number of different 'types') of smallwords used?

3 Does direct comparison of non-verbal filled pausing with smallword use, overall and in different turn-positions, yield evidence of an inverse correspondence between these features?

The way the variables listed above are measured, compared and assessed is discussed in detail in 'Findings on temporal variables' and 'Findings on smallwords' pages 165 – 177, which report on the actual analyses. However, a general introduction to the method is presented below.

Method

The analyses are based partly on data directly available through the corpus query form and partly on manually scanned printouts (pre-sorted according to student group) of whole transcripts or features listed in context. Figures for groups are compared, with the backing of chi-square testing.

In the case of temporal variables, the data has been scanned for filled pauses, which are defined as anything that might be described as a non-verbal utterance, such as *er, erm* and *uh.* The total occurrences of these were noted (ignoring the type of utterance) in the data for each group. The number of occurrences has also been noted for filled pauses in turn-initial, turn-internal (i.e. anywhere between the first and last word of the turn) and turn-final positions, as well as in 'sole' positions (i.e. where a filled pause is uttered by itself, and not accompanied by any words). The numbers of words and turns per group have also been noted, and the mean turn lengths (i.e. number of words per turn) calculated. Moreover, the ratios of words : filled pause and turns : filled pause have also been worked out. This information is all shown in Table 7.1 on page 167.

In the case of smallwords, a similar procedure has been followed to that for pauses. For each smallword, occurrences in context have been printed out, groupwise, and the numbers of occurrences noted, totally and in different turn positions. During this scanning procedure, smallwords have also been vetted for eligibility (see 'The smallwords' pages 162 – 163). The following turn-position categories are defined for smallwords: turn-initial, turn-internal, turn-final and 'loners' (i.e. where the turn is solely made up of one or more smallwords, alone or accompanied by *yes/no*). Moreover, the mean number of occurrences of smallwords per student has been computed as well as the proportion of the group who actually used individual smallwords at all. The total number of different types, i.e. the range of smallwords used by the groups, has also been noted. This data is tabulated and shown in full in Appendix I pages pages 291 – 292.

The sub-questions for questions 1 and 2 are addressed directly from the figures made available as outlined here. Question 3 is addressed by directly comparing the information yielded on pausing and smallword use overall and in different turn positions.

It should be mentioned that filled pauses have been included in total word counts, since I believe that they may sometimes be being used as 'alternatives' to smallwords, where the ability to use these is lacking. However, they have been ignored when assigning turn position to smallwords, so that, for example, *all right* in *erm, all right, I'll talk about ...* was regarded as being in turn-initial position.

Findings on temporal variables

This section covers the analysis of temporal variables in the data. The primary aim of this part of the analysis is to find independent evidence to support the claim (based on global grades) that the NoA group was more fluent than the NoB group. However, any evidence is noted that may lead to a way of describing the performance of students at different fluency levels with respect to these variables. Here we address the first of the research questions: is there evidence to indicate that the NoA group was more fluent than the NoB group, judged by:

- the occurrence of non-verbal filled pauses overall, and in different turn-positions?
- the mean length of turns?

These sub-questions are addressed on the basis of the raw data in Table 7.1 and the normalised data (per 10,000 words) in Table 7.2 (both tables are on page 167). Contingency tables are used to see whether the differences found between groups are significant (defined here as $p < .05$). It should be emphasised that the value of this data is limited in that it is too rough to enable absolute conclusions to be drawn – no account is taken, for instance, of how many of the total words are issued in running speech, and how many individually comprise whole turns. Findings are thus not regarded as more than indications, which are squared off with what common sense suggests can be expected.

The two variables referred to in the sub-questions are closely associated with what have been shown in 'Identifying elements of fluency' pages 126 – 132, to be widely recognised markers of fluency. The reason for opting to measure the mean length of turn, rather than the mean length of unbroken run, is discussed in 'Mean length of turn' on page 169.

The decision to measure the frequencies of non-verbal filled pauses only, i.e. excluding unfilled, silent pauses, has been made on the grounds that, although the transcripts were marked for the latter, these are not considered reliable; in the absence of any scientific means of measuring these, the

question of what is regarded as a pause is entirely subjective, and in the ear of the transcriber. Non-verbal filled pauses (e.g. *er* or *erm*) are more reliably identified as such, provided that no significance is attached to the individual way these are transcribed. In the present analysis no distinction is made between the different varieties of non-verbal filled pauses, the overwhelming majority of which occurred as *er*.

The legitimacy of restricting pausing to the filled kind in an analysis of this sort is addressed by a consideration of some of the studies previously cited, on the use of temporal variables in measuring fluency. Towell *et al.* (1996) restrict the use of the notion 'pause' to silent pauses only, defining the length of run as the number of syllables between pauses of 0.28 seconds and above. The very precision of this definition highlights the need for ultra-sensitive equipment to identify silent pauses, and gives justification to the policy in this study of abandoning any such attempt. Freed (1995), on the other hand, analyses the use of both filled and unfilled pauses, both of which she accepts as potentially being 'dysfluencies', or 'interruptions to the flow of speech' (1995: 131). Furthermore, she notes the idiosyncratic use of these two types of pause, one student marking herself out by using considerably fewer filled pauses and many more silent pauses than any other of the eight students studied. Lennon (1990) also analyses the frequency of both filled and unfilled pauses and concludes that the frequency of filled (but not unfilled) pauses per T-unit is one of the three variables emerging from his study as discriminating significantly between more and less fluent speakers, and should thus be included in the 'core of any set of measures for fluency assessment' (1990: 413).

Thus there is a precedent among studies seeking to identify temporal markers of fluency for recognising the filled pause as such a marker. Furthermore, any attempt to analyse unfilled pauses in the present dataset could yield highly unreliable results. And whether the study is undermined by the loss of data on silent pauses is, anyway, questionable. A scan through Fulcher's (1996) coded transcript example reveals that, out of 36 points in the dialogue coded for pausing of different types, 32 contained filled pauses (sometimes accompanied by a silent pause), three contained silent pauses only and one was coded on the basis of repetition only. This ratio of filled to entirely unfilled pauses, being ten to one, suggests that, in spoken interaction, pausing is normally filled, and so the question of whether it is justifiable to restrict the present analysis of pauses to that of non-verbal filled pauses may well be academic.

Filled pausing

Raw figures for the number of filled pauses used, as well as ratios between filled pauses and turns and words, are shown in Table 7.1, while Table 7.2 shows the number of filled pauses per 10,000 words.

Table 7.1 Group raw data on filled pauses, overall and with respect to position and number of words and turns

group	overall	turn-initial	turn-internal	turn-final	sole	total words	turns	mean words/turn	mean words/filled	mean turn/filled
NS	348	111	230	5	2	12349	707	17.5	35.5	2
NoA	845	134	674	17	17	14066	983	14.1	16.6	1.2
NoB	815	178	594	24	19	10467	1138	9.2	12.8	1.4

Table 7.2 Occurrences per 10,000 words of filled pauses, overall and with respect to position

group	overall	turn-initial	turn-internal	turn-final	sole
NS	282	90	186	4	2
NoA	601	95	479	12	12
NoB	779	170	567	23	18

As the tables show, the frequency of filled pauses in turn-final and 'sole' position was very low, so these have not been included in the analysis. Furthermore, what is identified here as a sole filled pause may be of the *hm* type, in which case it is probably, in fact, a backchannel rather than an actual pause. This section thus focuses on filled pausing overall and in internal and initial turn-positions.

When considering overall occurrences of filled pauses with respect to words, the pattern:

NS occurrences<NoA occurrences<NoB occurrences

was found to exist at a very high level of significance ($p<.0001$); in other words the 'better' the group (judged either by native-speaker status or by global grade), the fewer filled pauses were used. In fact, the native speakers used approximately one filled pause in every 35 words, the NoA group one in every 17 words, and the NoB group one in every 13 words. This pattern, NS<NoA<NoB, is found to be repeated significantly when comparing the filled pauses used in turn-internal positions. In initial position, analysis is carried out comparing pauses with respect to turns. The data for this analysis is shown in Table 7.3.

Table 7.3 Proportion of proper turns initiated by filled pauses

group	turn-initial filled pauses	proper turns	ratio pauses : turns
NS	111	651	1 : 5.9
NoA	134	891	1 : 6.6
NoB	178	1089	1 : 6.1

Here, only 'proper turns' have been counted, i.e. eliminating turns consisting only of 'loner' smallwords (see 'Method' pages 164 – 165) or sole pauses. This analysis reveals no significant difference between the groups, who all, in fact, exhibited remarkably similar behaviour: every sixth or seventh turn was initiated by a filled pause, as shown in Table 7.3 on page 167.

On the face of it, the result for overall filled pausing upholds the claim that the NoA group was more fluent than the NoB group. Moreover, the more detailed analysis has indicated that this difference in quantity of pausing was principally located in turn-internal positions. The evidence suggests that there was little difference between the pausing habits of the native speakers and the two Norwegian groups in getting started, but that the less fluent the speakers were assumed to be, the more they depended on filled pauses to keep themselves going. It must also be added that the NoA group was considerably more dependent on this kind of pausing than the native-speaker group.

At this point it is appropriate to raise the question of what kind of pausing seems to differentiate between the speaker groups. Fulcher (1993) identifies two types of filled pausing: 'content planning hesitation' and 'grammatical forward-planning hesitation'. He concludes that higher-ability students 'show a clear reduction in hesitation for grammatical planning and misunderstanding or communication breakdown and an increase in the use of [content-planning] hesitation to introduce examples or arguments' (1993: 151). Fulcher's conclusions may offer some underlying explanation of the findings on pausing in this study. While the less fluent students unquestionably paused more in general, this increase in pausing is clearly located within turns. In getting started, no difference in frequency of pausing has been noted between the groups. It seems reasonable to suppose that initial pauses – when a speaker is about to take the turn – are most likely to occur in planning content.

In mid-turn, there is no reason to assume that the NoB group needed more time to plan what to say than the more fluent NoA group. On the contrary, they probably had less need for this kind of planning; they produced shorter turns and, presumably, in line with Fulcher's findings (1993, 1996), they were less inclined to expand, e.g. through giving examples, on what they were saying. Because of their reduced language resources, however, it is fair to assume that they needed more time to plan how to say what they wanted. It can reasonably be assumed, therefore, that the increase in mid-turn pausing among the less fluent students can be interpreted as an increase in what will be referred to here as 'language-planning' pauses (i.e. pausing while trying to cope with language difficulties). Thus it can be concluded that the NoA group were more fluent speakers than the NoB group, judged by the difference in frequency of language-planning, and hence disruptive, filled pauses.

Mean length of turn

I have decided to consider the mean length of turn as a variable instead of the mean length of unbroken speech run. This is largely because the corpus does not allow for any accurate way of measuring the latter. However, it is also because I feel that the mean length of unbroken speech run is too closely related to the frequency of filled pausing, relative to total words, to provide a truly separate, non-circular, measure. Therefore I believe that the mean length of turn provides a preferable measure, because it is loosely associated with length of unbroken run, yet independent of the pause : word ratio, which has already been established as differentiating between the groups. This decision also reflects the contention made by Freed (1995) that a characteristic of more fluent speakers is 'more speech' (see'Forging a link between smallwords and fluency' pages 135 – 151). Since it is not possible in the present analysis to compare students according to the total amount of speech they produced, because tests varied considerably in length, the fairest way of comparing quantities of speech seems to be through considering the mean length of turn, measured in words.

A clear difference is found between the mean length of turns of the three groups. As can be seen from Table 7.1 page 167, the ratio of the mean lengths for the NS, NoA and NoB groups are 17.5, 14.1 and 9.2 words respectively. Thus, the NoA students – with relatively long mean turns – more closely resembled the native speakers in this respect than they did their weaker peers. This finding adds further support to the claim that the NoA students were more fluent than the NoB.

Conclusions on temporal variables

The measures of the two variables studied – disruptive pausing and mean length of turn – indicate that the NoB students were less fluent than the NoA students, who, in turn, were (not surprisingly) less fluent than the NS students. Moreover, certain other facts have emerged from the analysis which should be borne in mind when writing statements about fluent performance at different levels. The first is that, in line with what Fulcher (1993, 1996) maintains, pausing did not appear to function as a marker of student fluency universally; at the beginning of turns, it did not appear to differentiate between groups, only doing so when it occurred in running speech. The second is that even the more fluent Norwegian students paused considerably more than native speakers. However, these more fluent students managed to produce turns which, while rather shorter than the native speakers', were much longer than those of the less fluent students'.

Although there is no straightforward way of measuring the mean length of unbroken speech run, the joint findings of the decreased ratio of pausing to total words and the increased mean length of utterances of the NoA students

suggest that their mean length of run – bounded either by pausing or by turn boundaries – was longer than that of the NoB students. And following the reasoning of Towell *et al.* (1996), who found that the mean length of run was the greatest single contributor to a faster speech rate (see 'Identifying elements of fluency' pages 126 – 132), it can be assumed that the NoA students also produced speech faster than the NoB students.

To sum up, the NoA students have been found to be more fluent than the NoB students. They used relatively fewer pauses in mid-turn and were able to sustain longer turns in the dialogue, and hence can be argued to have produced longer stretches of unbroken speech and to have spoken at a faster overall rate.

Findings on smallwords

In the previous section it was corroborated, through comparing temporal variables, that the NoA students can be regarded as more fluent than the NoB students, but that even the NoA group appears to have fallen far short of native-speaker fluency. In this section, differences will be looked for in the quantity, distribution and range of smallwords used by the three groups. It is hoped that some patterns will emerge that will allow general statements to be written about smallword use at different fluency levels, and which will give rise to the need for a more detailed, follow-up analysis of how smallwords were used by the different groups. The sub-questions to question 2 posed in 'Hypotheses and research questions' pages 163 – 164, are addressed in turn; i.e. evidence of greater smallword use among the more fluent speakers is sought from:

- the total number of tokens, overall and in different turn positions
- the range (i.e. the number of different types) of smallwords used.

The complete dataset of frequencies for all smallwords analysed is shown in Appendix I pages 291 – 292, which contains the figures for each group and each smallword with respect to overall use and various turn positions, as well as the mean occurrences per student and the proportion of students who used individual smallwords. Corresponding figures are also given for the total of smallwords used, both as raw data and converted to ratios per 10,000 words. The range of different smallwords used by each group is also shown, overall and in turn positions. The analysis in this section uses data entirely taken from or derived from the dataset in Appendix I.

General smallword use: quantity and distribution

The raw data for tokens of general smallword use (i.e. irrespective of which smallwords are used), both overall and across different turn positions, is

summarised in Table 7.4. This table also shows the mean number of tokens per student, and the mean percentage of users of individual smallwords, as well as the range of smallwords, i.e. the number of different types produced.

Table 7.4 Group raw data on general smallword use with respect to overall use and distribution over turn positions

group	total	mean/ student	mean user percentage	range	turn-initial	turn-internal	turn-final	loner
NS	550	31	53%	19	185	302	10	54
NoA	393	21	32%	17	133	155	29	76
NoB	246	10	24%	15	110	84	20	28

The raw figures show quite unambiguously that the NS students were the biggest overall users of smallwords, with the NoA students comfortably in second place. The mean numbers of smallword tokens per student for the three groups, NS, NoA and NoB, were respectively 31:21:10. The mean proportions of students who actually used a particular smallword were one half of the native speakers, one third of the strong Norwegian group and one quarter of the weaker Norwegian group.

These facts appear to lend early support to the hypothesis that the more fluent Norwegian student group used smallwords in more nativelike quantities than did the less fluent group. It remains to be seen, however, if this evidence will stand up to more detailed, turn-position related analysis and to statistical testing.

As the student groups varied considerably in the total amount of speech produced (see Table 7.1 on page 167), the smallword data overall and in mid-term position are converted to relative data per 10,000 words, shown in Table 7.5.

Table 7.5 Group data per 10,000 words on general smallword use with respect to overall use and turn-internal position

group	total	turn-internal
NS	445	245
NoA	279	110
NoB	235	80

Contingency table testing on this data has revealed that, for overall smallword use relative to words, the relationship NS>NoA is found to hold with a very high degree of significance ($p < .0001$), with NoA>NoB significant at $p < .05$. In turn-internal position, this pattern is repeated. The mean proportions of 'proper turns', i.e. excluding loner smallwords (see 'Method'

pages 164 – 165), containing internal smallwords were approximately 50 per cent (NS), 20 per cent (NoA) and eight per cent (NoB). However, it must be remembered that the lengths of turns decreased dramatically from NS to NoB, so that these figures are not representative of actual smallword use in running speech.

In the case of smallword use in turn-extreme and loner positions, it was considered more important to carry out testing with respect to the number of turns uttered by the groups. For these non-internal turn positions, relative data per 1,000 turns is shown in Table 7.6.

Table 7.6 Group data per 1,000 turns on general smallword use with respect to non-internal turn positions

group	turn-initial	turn-final	loner
NS	261	14	76
NoA	135	29	77
NoB	97	18	25

For initial smallwords, the relationship NS>NoA>NoB is maintained with a high degree of significance ($p < .005$). In fact, the proportions of proper turns beginning with smallwords were roughly 30 per cent (NS), 15 per cent (NoA) and ten per cent (NoB). In turn-final position an unusual pattern emerged, with no significant difference between the two Norwegian groups, who both produced more smallwords than the native speakers. In the case of loners, the NoA group behaved very much like the NS group, while the NoB group produced very many fewer smallwords than the other groups ($p<.0001$). (However, it should be noted here that the NoB group, while using few loner smallwords, seemed to make up for this to some extent by an increased number of sole non-verbal utterances; see Table 7.2 on page 167. This use of non-verbal backchannels was noted by Fulcher (1996: 237) in his description of the performance of band 3 candidates.)

Thus it appears that the more fluent Norwegian students, while not using smallwords to the same extent as the native speakers, used significantly more than their less fluent counterparts, not only overall but in most turn positions. It seems that smallword use accompanied fluency, both in getting started in a turn and in keeping going. A striking result was that which showed that the more fluent students used loner smallwords, e.g. as backchannels, to a nativelike degree, while the less fluent marked themselves out by their low usage of these. This is in line with the findings of Fulcher (1993, 1996) and suggests that descriptors relating smallword use to fluency level should not only refer to the potential of smallwords to maintain the flow within a speaker's own turn, but also to the flow of the conversation across turns. The actual signals sent by smallwords used in this and other positions will be looked at in

the next chapter. In the meantime, it seems fair to conclude that more fluent speakers used smallwords to keep themselves and each other going.

Range and variety in smallword use

Table 7.4 on page 171, shows that the ranges, i.e. numbers of different types, of smallwords used by the NS, NoA and NoB groups were 19, 17 and 15 respectively. This suggests that the now-familiar pattern NS>NoA>NoB is also valid when it comes to the size of the 'pools' of smallwords the groups draw on. However, in order to see if this is a straightforward relationship, it is necessary to pose a few questions: Was the range of smallwords being added to, as the groups advanced in fluency, or did different groups use quite different pools of smallwords? Did the difference in range of smallwords apply in the various turn positions? And were the greater smallword ranges of native or more fluent speakers accounted for by odd occurrences of unusual smallwords, or by regular usage? This section will look into these questions in order to reveal whether the more fluent Norwegians used smallwords in a more nativelike way than the less fluent students, not only in quantities but also in the variety of their selections.

The first two questions can be answered by referring to Table 7.7 on page 174, which lists and ranks all smallwords that were used at least five times by groups in particular turn positions. The figure of five has been chosen arbitrarily, simply to exclude smallwords that were 'hardly ever' used.

In the case of the two turn positions where smallwords were most heavily used – initial and internal – the NS>NoA>NoB pattern can be detected by a glance at the lengths of the lists for the different groups. The superiority of the native speakers over both Norwegian groups is most visible in their range of turn-internal smallwords. On the whole, it seems that more fluent speakers have built onto the smallword vocabulary of less fluent speakers; all the words in the NoB lists appear among those in the NoA lists, which in turn are found in the NS lists, with the exception of *I think* in certain positions and *oh* as a loner. However, in terms of the ranking of smallwords according to frequency, there is no linear relationship across the groups. Top of the turn-initial list for the NS group is *right*, which does not even appear on the other groups' lists!

In turn-final position, where smallword frequencies were low, especially among the NS students, unsurprisingly, the learners showed slightly more variety in their selections. In the case of loners, only two types of smallword were used by each group, with *okay* being common to all groups, and again, the NS favourite, *right*, not appearing in the other groups' lists.

In order to shed light on the question of how the groups sampled from their pools of smallwords, a graphical depiction has been made of the extent to which different smallwords were favoured by the groups in turn-initial (Figure 7.1 on page 175), turn-internal (Figure 7.2 also on page 175), and loner positions

Table 7.7 Smallwords used in different turn positions, ranked in descending order of frequency

turn position	NS group	NoA group	NoB group
initial	right well okay all right ah oh just	okay well oh I think just ah	okay I think oh well
internal	just like well sort/kind of I think right okay or something a bit oh ah and things, etc.	I think just well or something okay a bit oh	I think just okay I think or something a bit
final	or something	or something I think	or something I think
loner	right okay	okay oh	okay oh

(Figure 7.3 on page 176). Pie charts are used so that the proportions of the different smallwords used by the groups are displayed. The pies themselves are roughly scaled in order to capture the differences in actual size of the body of smallwords used by different groups for any turn position. As these charts are intended to demonstrate the variety in selection of regularly used smallwords, all smallwords with tokens making up less than around five per cent of the total, or five tokens in all, are grouped together under the category OTHERS. The judgement regarding the variety of selection is based on the visual impact of the charts, which should reveal differences in the range of smallwords regularly used, as well as whether groups were relatively dependent on particular smallwords.

Figure 7.1 illustrates the way the different groups favoured smallwords when 'getting started' in a turn. In the case of the greatest users, the NS group, no smallword occupies a sector that is noticeably dominant in the chart. Five smallwords were used regularly. A clear preference was shown for *right* and *well*, but together these two smallwords only make up a little over half of the tokens produced. Neither Norwegian group used *right* at all and both showed

Figure 7.1 Smallwords used in turn-initial position by the three student groups

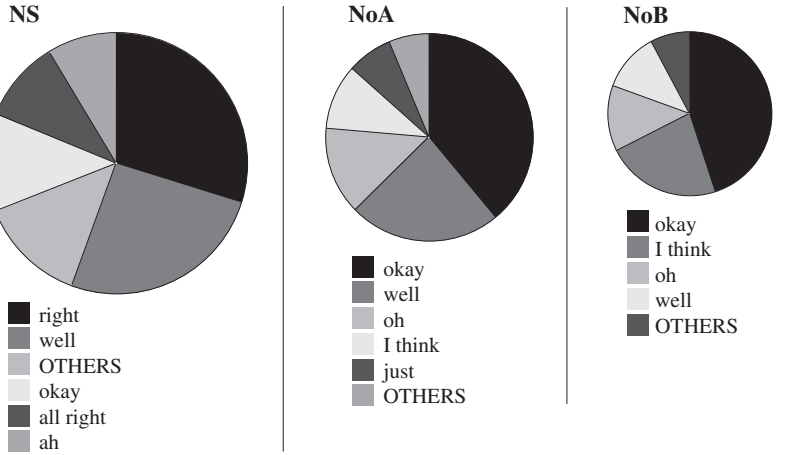
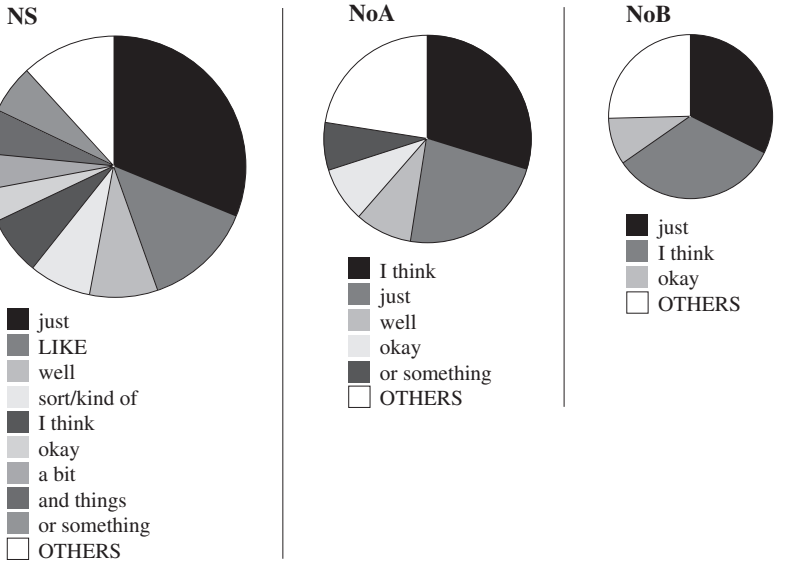


Figure 7.2 Smallwords used in turn-internal position by the three student groups

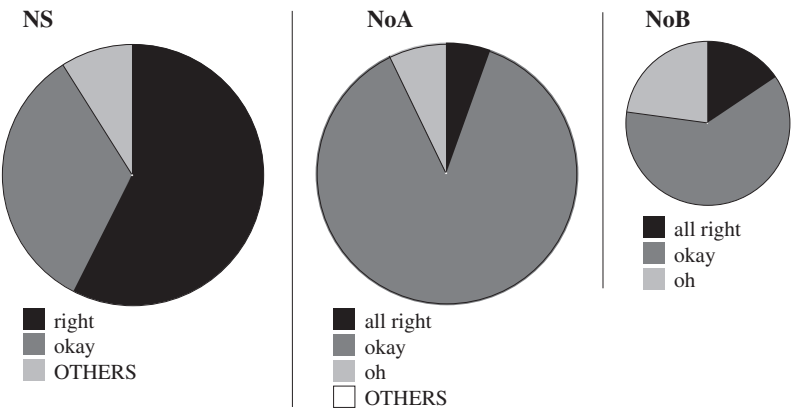


a distinct preference for starting with *okay*, which occupies the largest sector in both groups' pies, making up almost half the tokens. The NoA group used five smallwords regularly, with a nativelike placing of *well* in second place. The NoB group used four, with *I think* in second place.

Figure 7.2 on page 175 shows the groups' smallword preferences when 'keeping going' during a turn. Here, even for the least fluent group, no single smallword is seen to dominate for any of the groups, the 'favourite' smallwords not occupying much more than about a quarter of the area of the chart. The main feature that distinguishes the Norwegians from the native speakers is the NS groups' pool of nine regularly used smallwords, compared with five for NoA and three for NoB, whose graph is dominated by two smallwords – *just* and *I think*. *Just* was given a high ranking by all groups, but the Norwegians differed from the native speakers in their propensity for using *I think*, which was little used in this position by the NS group.

Figure 7.3 shows the ranges of smallwords used in loner position, where they can be thought of as being used to keep the conversation going across turns. While the NoA group produced a similar number of loner tokens to the NS group, it can be seen from Figure 7.3 that the groups differed markedly in their range of loner smallwords. While the NS group varied between two smallwords, favouring *right* ahead of *okay,* the NoA group overwhelmingly selected *okay*. The NoB students also favoured *okay*, insofar as they used loners at all, but these students also used significant proportions of *all right* and *oh*, which were used to a considerably lesser extent by the NoA students.

Figure 7.3 Smallwords used in loner position by the three student groups



Smallword use summed up

The findings from this analysis of smallword use seem to support the hypothesis that the more fluent students used smallwords in a more nativelike way than the less fluent ones did, as far as quantity and distribution across

turns are concerned. Overall, as well as in turn-internal position, the ratios of smallwords to words show that the NoA group used significantly fewer smallwords than the NS group, yet significantly more than the NoB group. This pattern was repeated when the ratios of turn-initial smallwords to turns were compared for the groups. Both Norwegian groups actually used more smallwords in turn-final position than did the native speakers. However, the NoA group showed themselves to use smallwords in loner position to the same degree as the NS group, while the NoB group seemed very little inclined to do this.

In their ranges of smallwords, further striking differences have been revealed between both Norwegian groups and the native speakers. While the more fluent Norwegian students had a rather wider range of smallwords in regular use than the less fluent students did, they generally used a narrower range than the native speakers and were more inclined to let certain smallwords dominate in a turn position. The smallwords used by the Norwegian groups were, however, among those generally used by the native speakers, but with some exceptions. The NS favourite, *right,* in non-internal position was not used by the Norwegians, who depended heavily on *okay*. In turn-internal positions the Norwegians used *I think* to an extent that was atypical of the native speakers.

The findings on the ranges of smallwords used by the groups echo those of other researchers. Raupach (1984) found that, even after a stay abroad, having acquired extended vocabularies, students as a group stuck to a very similar and restricted repertoire of organising formulae (see 'Identifying elements of fluency' pages 126 – 132). And Nikula (1996), in her study of 'hedge-like' modifiers among Finnish speakers of English, says: 'As far as the types of expression used are concerned, the non-native speakers had a narrower range at their disposal even though they used most of the modifiers that ranked highest in the native speakers' performance' (1996: 90).

In short, the more fluent Norwegian students seem to have been more nativelike than the less fluent students in the extent to which they used smallwords in getting started and in keeping themselves going in their turns. Moreover, they appear to have used loner smallwords to the same extent as the native speakers in keeping their partner's side of the conversation going. However, even in this last respect, they fell short of the native speakers in the size of the pool of smallwords they drew on. It seems, in other words, that smallwords may be an area of vocabulary where learners most need the comfort of familiar words and phrases – the 'lexical teddy bears' of speaking (see Hasselgren 1993).

Smallwords and filled pausing

The third and last of the research questions to be investigated here read, in 'Hypotheses and research questions' pages 163 – 164, as follows:

> Does direct comparison of pausing with smallword use, overall and in different turn-positions, yield evidence of an inverse correspondence between these features?

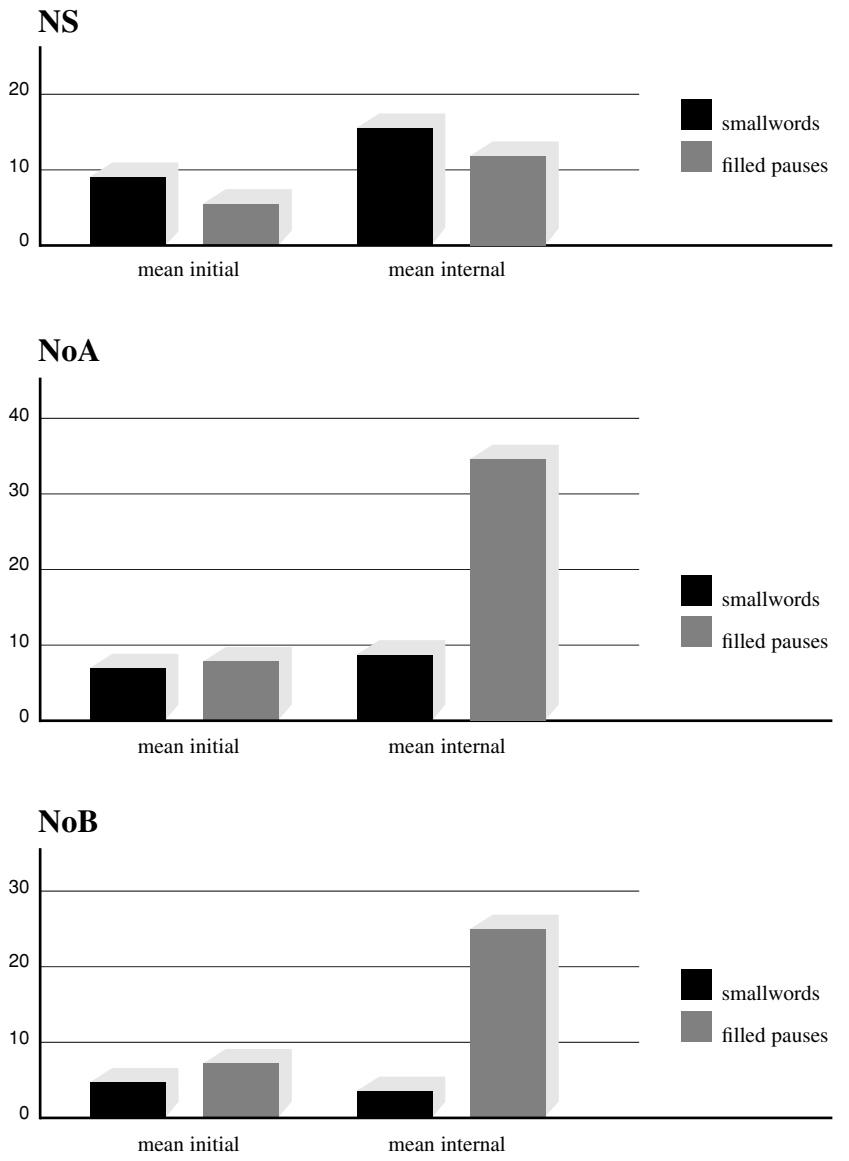
Put simply: did students who used more smallwords use fewer *erms, ers* and *ums*?

It has already been established, in this and the preceding two chapters, that there is more to fluency than a lessening of pausing. Moreover, the contribution of filled pausing to fluent speech should not be underestimated. Clark (1996) acknowledges that *suspension points*, at which 'speakers cease their presentation [...] where, in the ideal delivery, they should not be ceasing' (1996: 259) are normal to utterances. He further states: 'the hiatus in fluent speech – the interval between the suspension and the resumption point – is often filled with more than a silence' (1996: 262), going on to cite six common types of hiatus 'content', which include both 'fillers' (e.g. *uh, um*) and 'editing expressions' (such as *I mean*). In other words, fluent speech depends on both filled pauses *and* smallwords, and it would be crass to suggest that either one should be fostered to the exclusion of the other.

However, the fact is that, as language teachers are only too well aware, and as research, such as that of Lennon (1990), has supported, less fluent speakers use filled pauses excessively. Moreover, it has been demonstrated here that the excesses in filled pausing of the less fluent speakers are located in turn-internal positions, appearing to be associated with language planning, and can be considered disruptive. It seems logical to deduce, from the findings in this chapter so far, that it is this excessive, disruptive filled pausing that smallword use can counteract. In order, therefore, to further corroborate the general claim made here that smallwords play a part in contributing to fluency, as well as to illustrate the direct impact of smallword use on speech, a comparison is made of the smallword use : filled pausing ratios in the speech of the three student groups.

Contingency table testing has been carried out, comparing the frequency of filled pauses with respect to the number of smallwords across the groups in turn-internal and turn-initial positions. In turn-internal position, the relationship NS>NoA>NoB holds at a highly significant level ($p<.001$) for the ratio of smallwords to filled pauses (in other words, the more fluent the speaker group, the more often a smallword was used in mid-turn rather than a filled pause). In turn-initial position a similar result is found, although at a slightly lower degree of significance ($p<.01$). In the latter position, this difference is predictable and obviously due to the differences in frequency of

Figure 7.4 Mean numbers of smallwords and filled pauses per head compared for all three speaker groups



smallwords, since it has already been established that turn-initial pausing did not vary significantly between the three groups.

The filled pause-smallword link is rather simplistically illustrated in Figure 7.4 on page 179, which shows, for each of the three speaker groups, the mean number per head (using raw data) of filled pauses and smallwords respectively in turn-initial and turn-internal positions. When interpreting these charts, it must be borne in mind that the native speakers used rather fewer turns each than the Norwegians, and that the length of turn significantly increased in the order NoB<NoA<NS. This means that comparison of the heights of the columns across the groups is not viable, and that comparisons must be confined to column heights within the groups.

The NS group is seen to differ from both Norwegian groups mainly by the fact that, in both turn positions, the native speakers used more smallwords than filled pauses, using rather more of both of these in turn-internal than in turn-initial position. The Norwegian groups, on the other hand, while using slightly more filled pauses than smallwords in getting started (particularly in the case of the weaker group) used vastly more filled pauses than smallwords once they were under way.

In conclusion, the results of this analysis highlight a tangible way in which smallword use appears to affect speaking. The relationship between smallwords and pausing at the start of turns remains a fuzzy one (the more fluent groups used more smallwords but not fewer filled pauses). However, there is clear evidence to suggest that, in mid-flow, the tendency is for more fluent speakers to use smallwords at the expense of filled pauses, with the less fluent speakers doing the reverse. In fact, in turn-internal position, the NoA group produced four filled pauses to every smallword while the ratio for the NoB group was seven to one. The native speakers, on the other hand, actually produced slightly more smallwords than filled pauses in this position, in the ratio 1.3 to one!

It may be that smallwords used at the beginning of turns play roles in bringing about fluency that have no effect on initial filled pausing (which is probably mainly needed to collect ideas); these roles may, for instance, relate to across-turn cohesion, and will be commented on in the next chapter, which deals with the actual signals sent by smallwords. However, there seems little doubt that the excessive, mid-flow, disruptive filled pausing of the less fluent speakers goes hand in hand with their low smallword use, which adds substance to the belief that fluency, somehow, depends on smallwords.

Summary

This chapter has addressed a number of questions, the most central being whether more nativelike quantities and distribution of smallwords appear to go hand in hand with more fluent speech. Preliminary to tackling this question,

another issue has had to be addressed: temporal variables, established as markers of fluency in Chapter 6, have been used to check that the grouping of students into more and less fluent, on the strength of global grades, is valid when judged by independent measures. Moreover, since the overall aim of the investigation has been to shed light on the behaviour of fluent and less fluent students, with the writing of descriptors in mind, any information that could add to the understanding of this behaviour has been looked for.

The first part of the investigation has corroborated the claim that the NoA students, as a group, were more fluent than the NoB students, judged by the temporal variables of filled pausing and mean length of turn, and by the length of unbroken speech runs and the speech rate. It has also been revealed that, in the case of the frequency of filled pausing, only when this was measured in turn-internal position did it actually differentiate between the students at different levels of fluency and between the natives and non-natives. In start-position in a turn, filled pausing was found to occur in fairly equal quantities across the groups.

The second part – looking into smallword use – has shown clearly that, while not matching native-speaker use, the more fluent Norwegian students used smallwords in greater and hence more nativelike quantities than did the less fluent students, both overall and in most turn positions. Using them as 'loners', the stronger students actually matched up to the native speakers' performance in sheer numbers. In terms of variety or range of smallwords used, however, the more fluent students are shown to have been lagging far behind the native speakers, although emerging as somewhat wider users of different smallword types than their less fluent counterparts.

The third part of the investigation has drawn together the first two parts in maintaining that an increase in smallword use should, logically, accompany a reduction in filled pausing. This has been found to be the case, most notably in turn-internal position, showing that the excessive *erms* and *ers* of weaker students went hand in hand with a paucity of smallwords.

So far in this conclusion, care has been taken not to make claims that causes of fluency have been unearthed. This is because the findings, as they stand, cannot do more than confirm associations or correlations between certain features in the students' speaking. It could in fact be argued that more fluent speakers, in producing longer turns, were more likely to need and therefore produce smallwords. It may also be argued that the more frequent filled pausing of weaker students was due to their general language deficiencies, and would have happened regardless of smallword prowess. Both of these claims probably contain an element of truth; no attempt will be made here to prove or disprove them.

However, the findings here must be interpreted in the wider framework of the study. The theoretical argument was put forward, in Chapter 6, that smallwords perform tasks that directly contribute to fluency as it has been

defined here. Moreover, the empirical findings of other researchers, such as Towell *et al.* (1996), indicate that learners, having equipped themselves (e.g. during a stay abroad) with formulae of varying kinds (including smallwords), are able to produce speech with all the hallmarks of fluency, as they could not before, when they lacked these formulae. Therefore, while no claim is made here that smallwords are a panacea for dysfluency (which is caused by all sorts of things), it seems reasonable to claim that the findings support the overriding causal hypothesis that smallwords make a significant contribution to fluent speech.

Additionally, the findings themselves have, I hope, made a contribution to what is known about the behaviour that characterises students at different levels of fluency. A more tangible basis is emerging for describing this behaviour in terms of both temporal variables (i.e. the relative frequency of disruptive, turn-internal filled pausing, the length of turn, and, by implication, the length of unbroken speech runs and the rate of speaking) and smallwords, used overall and in various positions. This means that the writing of revised descriptors can draw on information beyond that normally associated with fluency indicators; this has valuable consequences for the washback validity of the test as well as for its reliability. Moreover, the ultimate description of fluency need not be confined to references to the general flow of a person's own speech, as in the case of temporal markers. Besides taking account of the finer details of getting started and keeping going, reference can also be made to the ability to contribute to the flow of the conversation itself, by supporting and acknowledging the speaking partner.

Before considering how these findings might influence the revision of fluency descriptors, however, it is necessary to perform another major analysis, that of determining how students use their smallwords to promote fluency. This analysis is the subject of the next chapter.

8 The signalling power of smallwords

It has become clear from the previous chapters that the analysis of smallwords is mainly driven by the quest for information that can be used in building up better descriptors of fluent performance. Moreover, this information should preferably be of a linguistic nature, i.e. to do with actual language features typical of students at different levels of fluency. Obtaining this information may be regarded as the primary, immediate goal of the analyses. However, it would be naïve to imagine that any linguistic feature that happens to discriminate between high- and low-fluency performances should automatically qualify for inclusion in descriptors. Over-complex descriptors can be unusable, and over focusing on details – such as the detailed use of particular smallwords – could be at the expense of valuable higher-order features – such as general vocabulary range.

The question then arises as to whether we already have enough information on the students' use of smallwords – in terms of quantity, turn position and variety – to enhance fluency descriptors sufficiently. The analysis covered by this chapter is justified on the grounds that the answer to this question is *no*, for three reasons. Firstly, although it has been established, in terms of numbers, that the more fluent Norwegian students (the NoA group) used more smallwords than the less fluent (the NoB group), we need to know the extent to which they were used appropriately – i.e. in the way native speaker students used them. The 'use' reported on here may in fact be 'misuse'. Secondly, we have as yet no idea of how students' underuse of smallwords actually affects performance. We need to know, for example, if it leads to a weakening of potential smallword signals across the board, or whether some signals are out of range for the speaker group, while others are fully supplied. Even quite general band-scale and performance profile statements about smallword use are dependent on these types of information. And thirdly, in the interests of the consequential aspect of validity of the test, if descriptors reveal the need for greater competence in smallword use, teachers need to be equipped to help build up this competence. It is hoped that the level of detail in the present analysis will contribute substantially to the knowledge both of how native speakers generally use smallwords and of the order in which learners apparently acquire this usage.

These three needs give rise to the three aims of the analysis reported in this chapter. The first is to find out whether the Norwegian groups used their

smallwords in a reasonably nativelike way. The second is to find out the broader effects of the Norwegian students' smallword underuse on their fluency in performance. And the third is to build up a greater fund of detailed knowledge about the order in which smallword use appears to be acquired by learners.

The approach

The analysis to be covered in this chapter is loosely defined as 'qualitative'. This is in contrast to that covered by the last chapter, where quantities were paramount, and conclusions were made largely on the basis of statistical testing. In Chapter 6, it was claimed that smallwords contribute to fluency by sending certain microsignals, shown in the framework in Figure 6.2 on page 155 (and referred to simply as 'signals' from now on). This framework forms the structural basis for the analysis in this chapter. Each signal is examined in turn, for evidence that smallwords are being used to send it. As a preliminary step, each occurrence of a smallword in the dataset has been examined with respect to the signal(s) that it appears to be sending. Evidence that smallwords are sending particular signals is primarily based on the context in which the smallword occurs, but also takes into account what other researchers have said about it, and, in the case of certain smallwords, any explicit inherent meaning it may have. Numbers are considered only in assessing the relative weight of evidence that smallwords appear to be used to send particular signals.

This approach differs somewhat from that of many valuable pieces of research, e.g. those reported in Heritage (1984), Stenström (1984), Schiffrin (1987) and Jucker (1993), on the functions of words or phrases ('markers') used by native speakers. In these studies the researchers have applied their intuition about the marker itself to what the data reveals about its contexts of use, drawing conclusions on the use to which the speakers seem to be putting the marker. Such an approach is in line with Schiffrin's (1987) claim that what a marker contributes to the discourse is a function of its inherent 'meaning' and its location with respect to the various planes of discourse (1987: 63). However, these studies have involved more or less homogeneous groups of speakers, whose usage is presumably similar to the researcher's. By contrast, in the present study of the language of learners, few assumptions can be made about the meanings that individual speakers attribute to most smallwords (although there are a few signals, notably hedging and appealing, which are given inherently by smallwords themselves).

In this analysis, therefore, with a few exceptions, smallwords are not normally predefined as actually giving particular signals. Instead, typical 'contextual slots' are defined for each signal, these being the points in the discourse where the signal seems appropriate. The number of tokens of smallwords occurring in these contextual slots are noted for each student group.

If it transpires that a particular smallword, or group of smallwords dominates in a particular contextual slot, this will be taken as evidence that the group is using the smallword(s) to give that signal. Example [1] can be used to illustrate this:

[1] T: ... just say what's happening
A: well , I think they are , these two are , they're going to see a match and , **well** , eh , they are now in the , in a station café and , eh , (44:18) (NS)

Here the (second) *well* occurs at a point where the speaker interrupts his/her own flow and embarks on a new formulation of the message, i.e. in the contextual slot for a mid-utterance break (see 'Signalling a mid-utterance break with context created by the speaker's own immediately preceding speech' pages 196 – 200). This individual occurrence of *well* will not be taken, *per se*, to be signalling a mid-utterance break. However, if many tokens of *well* are found to have recurred in this contextual slot, the group (in this case the NS student group) will be assumed to have, apparently, been using *well* to give this signal.

Where the NS students seem to have been using a smallword as a signal, the literature is consulted in order to reinforce the impression that this is normal usage of the smallword. This is done partly because the numbers of tokens of smallwords is sometimes small, and to counteract any highly (locally) idiosyncratic usage. Having established which smallword(s) the NS students appear to have been using (normally) to give a signal, the Norwegian groups are assessed with regard to whether they appear to have been using <u>any</u> smallwords to give that signal, and if so, whether these were the <u>same</u> smallwords as those used by the NS group.

It must be emphasised that 'contextual' does not necessarily refer here to the immediate textual context. The 'discourse slot', which Schiffrin (1987) claims to be an essential factor in determining a marker's function (1987: 73), can be on any plane of discourse, such as *participational, ideational, action*wise and *exchange*wise (see 'Smallwords in other peoples' books' pages 135 – 138). This multiple-plane concept is preserved in the present study, where contextual slots for signals have been defined according to quite distinct contextual planes, such as the turn structure, the task that is being carried out, the idea that is being expressed by the speaker, or what the interlocutor has just 'done'.

Data, hypotheses and research questions

The data analysed is identical to that used in Chapter 7. The examples of smallword use have been obtained from the EVA Corpus and printed out in

context (with approximately ten words on either side) and sorted according to the three student groups, NS, NoA and NoB.

Evidence to suggest that smallwords are being used to give a signal is based on either the recurrence of the smallword in the contextual slot for the signal or, in certain cases, the explicit, inherent meaning of the smallword itself.

Two main hypotheses are tested in this investigation:

- the more fluent students will be found to have used smallwords to send a greater range of signals than the less fluent students, and to have sent these signals with a more nativelike range of smallwords
- both Norwegian groups will be found to have had gaps and limitations in the signals they sent with smallwords.

In order to test these hypotheses, the following research questions are posed for each of the signals identified in 'A framework for analysing smallword signals' pages 151 – 155:

- Is there evidence to suggest that the groups used smallwords at all to give this signal?
- Is there evidence to suggest that particular smallwords appear to have been used by the groups to give this signal?
- Is there secondary evidence (in the literature) that the smallwords favoured by the NS group to give this signal are acknowledged as normal native-speaker usage?
- Did either of the Norwegian groups appear to use any or all of the smallwords that the NS group apparently used to give this signal?

Method

A preliminary stage of the analysis has involved deciding what should constitute evidence that smallwords send signals. In the case of some signals (i.e. hedges and, to some extent, acknowledgers and appealers), where the signal is most easily identified by the smallword itself, this has involved deciding which smallwords may be inherently regarded as sending the signal (e.g. *sort of* and *you know*). However, in all other cases, it has meant defining the contextual slot(s) for the signal. Defining contextual slots has been a painstaking process, involving many rounds of trial coding by myself and a second native speaker. These definitions have had to be tight enough for both to agree on codes, and broad enough so as not to exclude many cases that both intuitively felt should be included.

To further clarify what is meant by a contextual slot, and how tokens occurring in such a slot are classed as evidence, the example of signalling a mid-utterance break is briefly reconsidered. A detailed account of this

signalling is given in 'Signalling a mid-utterance break with context created by the speaker's own immediately preceding speech' pages 196 – 200, but it suffices here to say that mid-utterance breaks occur when a speaker breaks into his/her flow with some sort of comment or repair, or to return to a previous theme. This occurs in examples [1] (above) and [2] (below):

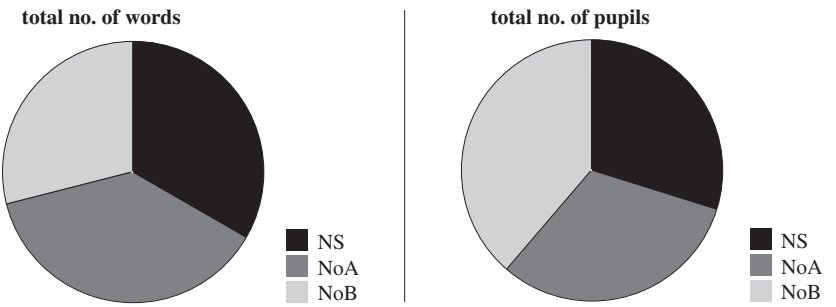
[2] A: all right . well , they've gone to the café , and then they go to buy tickets . **oh** no , he asks her where the tickets are , and she says ... (49:8) (NS)

The typical contextual slot for a mid-utterance break signal occurs between what is first said and the new or 'repaired' part of the utterance. As exemplified in examples [1] and [2], taken from the NS data, both *well* and *oh* recurred in this slot (5 *well*s and 7 *oh*s out of a total of 15 smallwords). This evidence suggests that the NS students used both *well* and *oh* to give this signal.

The main body of the work has involved the final coding of each smallword, principally according to the contextual slot it occurs in. Tapes have been listened to, where it was not possible to agree on codes from the transcripts alone, e.g. due to the absence of intonation marking. Appendix J pages 291 – 292, presents a numeric overview of the signalling functions assigned to smallwords by the three student groups.

The analysis concludes by assessing which signals seem to have been sent through smallwords at all by the Norwegian students, and whether this apparent signalling was done using some or all of the range of smallwords normally used by native speakers. Although tallies are low, conclusions are drawn on the assumption that, all things being equal, the Norwegian groups should have had roughly the same number of occasions for sending a signal as the NS students. It should also be borne in mind that the numbers of words produced by the student groups NS/NoA/NoB were in the rough ratio 12.5 : 14 : 10.5, while the numbers of students (and hence the number of tasks done) were in the ratio 18 : 19 : 24. These ratios are shown graphically in Figure 8.1.

Figure 8.1 Group proportions of the total numbers of words and students



Since there are no huge discrepancies in the groups' proportions of words or students, a striking difference in token numbers is interpreted as an indication of a difference in usage. The gaps or limits in signalling that are thus indicated offer some kind of explanation of how dysfluency occurred in the two Norwegian groups.

Defining and analysing evidence that smallwords are used to send signals

The definitions, analyses and findings are discussed for each signal in turn in the following sub-sections. First, the actual definitions of contextual slots for each signal are presented. Next, the number of tokens of smallwords occurring in this slot, (or being associated with the signal by virtue of inherent meaning) is tabulated and compared for each group. Reference is then made to the literature relevant to any smallwords that the NS students seemed to associate with the signal. Finally, we assess the way the Norwegian groups compared with the native speakers in their use, or non-use, of smallwords to give this signal.

The signals analysed are those identified in the framework shown in Figure 6.2 on page 155. The definitions and analyses of evidence are grouped according to the five macrosignals laid down in that framework. For an initial definition of each signal, the reader should refer to 'A framework for analysing smallword signals' pages 151 – 155. The five macrosignals under which signals are classified are:

1 EXPRESSING THE COMMUNICATIVE INTENTION
2 POINTING TO THE CONTEXT FOR INTERPRETATION
3 INDICATING THE COGNITIVE EFFECT OF THE PREVIOUS UTTERANCE
4 INDICATING THE DEGREE OF VAGUENESS OR COMMITMENT
5 INDICATING THE STATE OF 'SUCCESS' OF COMMUNICATION.

Expressing the communicative intention

Two signals have been identified, on pages 151 – 155, as making up the macrosignal 'expressing communicative intention'. The first involves the signalling of the intention to take, hold or yield the turn. The second involves signalling that the speaker does not intend to give the response that the previous speaker might have expected or hoped for; in other words, an 'oblique response' is signalled.

Signalling whether the speaker intends to take, hold or yield the turn

DEFINING EVIDENCE

The contextual slot for signalling whether the speaker intends to take, hold or yield the turn is defined simply in terms of the position within the turn structure (see 'Method' pages 164 – 165). For example, turn-initial position is (rather obviously) regarded as the slot for signalling that the speaker wishes to take the turn, occupied by *okay* in [3]:

[3] T: so , you know look at the the directions and then tell her
A: **okay** er from the boat to Newcastle er Central Station there's er a bus , you can take that , and ....
(02:251-2) (NoA)

ANALYSING EVIDENCE

The way students use smallwords to signal their intention to take, hold or yield the turn is based on the evidence of the across-turn distribution of smallwords, already discussed at some length in 'General smallword use: quantity and distribution' pages 170 – 173. Where the numbers of tokens indicate that smallwords were typically used in a certain turn position, it is concluded that they were being used by the group to signal turn-taking, turn-holding or turn-yielding, according to the turn position.

The findings on pages 170 – 173 provide ample evidence that all groups used smallwords to signal that they were to take or hold the turn. The relationship NS>NoA>NoB has been found to hold for the quantities of both of these signals, implying that, while not matching the native speakers, the more fluent Norwegian students were more likely to send these signals using smallwords than the less fluent students were. There is, on the other hand, little evidence to suggest that the NS students signalled a yielding of the turn through smallwords. While the NoA groups actually behaved in a less nativelike way than the NoB group in this respect, it must be concluded from the very low overall numbers of turn-final smallwords that the Norwegian students resembled the NS group in that they did not habitually signal turn-yielding through smallword use.

Insight into which smallwords the groups appear to have used as signals of turn-taking and -holding can be gained from Figures 7.1 to 7.3 pages 175 – 176. These figures show clear differences in the preferences of the groups as to which smallwords they used in turn-taking and turn-holding slots. Table 8.1 on page 190 (drawing on the full dataset in Appendix J pages 293 – 294) shows the raw data for the most significantly used (i.e. more than five times by any group) smallwords in the contextual slots for signalling turn-taking and turn-holding:

Table 8.1 Raw data on smallwords occurring in contextual slots for signalling turn-taking and turn-holding

	turn-taking			turn-holding		
smallword	**NS**	**NoA**	**NoB**	**NS**	**NoA**	**NoB**
right	55	-	-	14	1	-
all right	18	2	4	1	2	-
okay	22	52	51	11	14	9
well	48	32	12	24	15	4
oh	6	17	15	8	6	4
ah	15	4	1	5	1	1
I think	7	16	24	22	46	26
just	5	7	2	96	35	27
kind/sort of	1	1	-	23	4	-
like	3	-	-	43	3	1
a bit	1	-	1	16	7	5
or something	-	-	-	12	11	5
and things, etc.	-	-	-	12	2	2
others	4	2	-	15	7	-
total	**185**	**133**	**110**	**302**	**154**	**84**

It can be concluded from the data shown, together with the discussion in Section 8.3.2, that the NS students favoured *right* as a signal of turn-taking, alternating this with both *all right* and *okay* and using *well* when the context demanded it (as will be discussed in 'Pointing to the context for interpretation' pages 194 – 200, for example). The Norwegian students, on the other hand, tended to signal turn-taking overwhelmingly with *okay*, although the figures for the NoA students indicate their awareness of the potential of *well* as a turn-taking signal.

This clear preference among native speakers for using *right* to signal turn-taking is reflected in a study of academic adults' use of *right, all right* and *okay,* in the London-Lund Corpus, by Stenström (1990: 165), who found that *right* made up 50 per cent of the turn-initial uses of these words.

The smallwords used to signal turn-holding might be regarded as belonging to two categories. The first category consists of explicit hedges (whose use is analysed in more detail in 'Indicating the degree of vagueness or commitment: Signalling a softening of the impact of the message, or "hedging"' pages 204 – 213), such as *I think* and *just*, which are depicted in the lower part of Table 8.1 (from *I think* downwards), while the other category, called 'non-hedges' for convenience here (although these may sometimes function as hedges) is depicted in the upper part of the table. For all student groups, the hedges made up somewhere in the region of 80 per cent

of all the smallwords used in this position, and it is in their range of hedges that Norwegian students differed most noticeably from the NS students. Hedges are discussed at some length in 'Indicating the degree of vagueness or commitment: Signalling a softening of the impact of the message, or "hedging"' pages 204 – 213, and therefore are not expanded on here.

Among the non-hedges used to signal turn-holding, all groups used *okay* with more or less the same relative frequency (around ten times per 10,000 words). The NS group used *right,* with a similar frequency, and *well* approximately twice as often. Neither Norwegian group used *right*, and only the NoA group used *well* to a significant extent (around ten times per 10,000 words).

What is also worth noting about this group of non-hedge smallwords in internal position is that more than half of them, in the case of the NS group, occurred in semi-initial position (defined here as following another smallword, or *yes* or *no*), i.e. in 'slot 2' in an utterance, using Stenström's (1990) terminology. This implies that they were, in fact, used, following some kind of acknowledgement, to signal that the speaker intended to take the turn, and thus they can be considered closer to turn-taking than to true turn-holding signals. Having ready access to a variety of smallwords might allow a speaker greater freedom to combine two or more smallwords when taking the turn. The NS group provided 27 turn-semi-initial smallwords, compared with ten (NoA) and three (NoB). This demonstrates that the combining of smallwords in signalling turn-taking is a nativelike feature which only the more fluent Norwegian group went some way towards emulating.

It must be concluded from what has emerged in this section that, while the Norwegian students cannot be said to have had 'gaps' in this signalling, they were less inclined to use smallwords to signal turn-taking and turn-holding than were the NS students. The NoA group did, however, perform in a more nativelike way than the NoB group in this respect, sometimes using combinations of smallwords. The NoA students were also more likely to make use of *well* in signalling turn-taking, but failed, remarkably, to use the signal *right* – most commonly used by the NS group to send this signal. In signalling turn-holding, apart from a significantly greater use of *well,* the NoA students were essentially no more nativelike than the NoB students in their range of smallwords used. This is most strikingly seen in their paucity of types of hedges (which made up the overwhelming majority of turn-holding signallers for all groups). As most of the smallwords concerned in signalling turn-taking or turn-holding also send other signals simultaneously, more will be revealed about the extent and implications of these limitations in the forthcoming sections.

Signalling an oblique response

DEFINING EVIDENCE

When a speaker's response does not (to a greater or lesser extent) communicate what the previous speaker seems to expect, it is referred to here as an 'oblique response'. The point immediately prior to either of two types of response can be considered the typical contextual slot for signalling an oblique response. The first type is recognised as when the proposition expressed by the speaker is not the response that the previous speaker can be presumed to be expecting. Such responses are referred to here as 'dispreferred responses', following Levinson (1983: 307) and Yule (1996: 79). These are identified as when the second speaker does not comply with what the first speaker can be assumed to be hoping for, e.g. in not granting a request, as in [4], where *well* occurs in the contextual slot for this signal:

[4] B: could I speak to your father please
A: **well** , he's not at home ... (02: 281 – 282) (NoA)

The second form of oblique response occurs when the speaker is unable or unwilling to give an immediate and direct response to the preceding utterance. This is recognised by the markers of ambiguity, e.g. *yes and no,* or markers of a 'qualified' or even contradictory response, such as *but, if* and *it depends,* as in [5] and [6], or a delaying of the actual answer with some kind of an appeal for help or clarification, as in [7]. Again, in all of these cases, *well* occupies the contextual slot for the signal:

[5] T: what about the first one . it says there the parents should decide when a teenager or a fifteen-year-old comes home at night . do you agree with that or
A: **well** , yeah , I agree if it's not unreasonable
T: mhm , what what would you say was reasonable
A: em , **well** , it depends where she was going ... (41: 53 – 56) (NS)

[6] T: ... could you say why could you say a bit more about that
B: eh **well** there is a lot of horror stories going around about hitch hiking but in most situations if you get someone to hitch-hike with <unclear> it is usually people that have well intended people (14: 68 – 69) (NoA)

[7] T: and she has to tell you where that is
A: **well** ... I don't even know what it is in English (02: 237 – 238) (NoA)

ANALYSING EVIDENCE

Only one smallword, *well,* is found in the contextual slot for signalling an oblique response to the previous utterance. Both the NS and the NoA students used *well* in this context (ten and 13 times respectively), and so can be regarded as having used *well* to signal an oblique response, while the NoB group, with only two tokens, cannot be considered to have done so. Table 8.2 shows the raw data for occurrences of *well* in the contextual slot for signalling an oblique response. The figures for NS and NoA tokens suggest that both groups produced tokens at roughly the same frequency, relative to the number of words uttered by each group.

Table 8.2 Raw data on tokens of well in the contextual slots for signalling an oblique response

smallword	**NS**	**NoA**	**NoB**
well	10	13	2

The uses of *well* cited here reflect what has been well documented from research into native-speaker speech. R. Lakoff (1973) sums up the use of *well* by stating: '*well* is used in case the speaker senses some sort of insufficiency in his answer, whether because he is leaving it to the questioner to fill in information on his own or because he is about to give additional information himself' (1973: 463).

Lakoff's point is borne out by the findings of Stenström (1984: 183). Jucker (1993) expands on Lakoff's point by showing that it is a discrepancy of the assumptions of the speakers that brings about 'some sort of insufficiency' (1993: 442) in the response, which is then marked by *well.* These discrepancies can be seen here as in assumptions about the linguistic ability of speaker A in [7], A's ability to give an opinion in [5], or A's ability to comply with a request as in [4]. Jucker sums up the use of *well* in answering questions thus: 'Answers that fail to supply the information required by the question are habitually introduced by *well'* (1993: 442). This summary can be taken to account for the use of *well* in all the data examples cited here.

Thus it may be assumed that both the NS group's, and the NoA group's use of *well,* in signalling an oblique response, is in line with what has been observed in the literature on this function of *well.* Both groups appear to use it to signal a response which is not quite what the other speaker seemed to be expecting, thus preparing the other speaker for some kind of adjustment or renegotiation. The relative weight of evidence suggests that, as a group, the NoA students use a smallword to signal the communicative intention to give an oblique response in a fully nativelike way. The NoB students, on the other hand, do not signal this response with any smallword.

Pointing to the context for interpretation

In 'The work of smallwords in optimalising fluency' pages 142 – 148, it was established that the normal, or default, context for interpreting an utterance is the 'initial context', i.e. what has just been said by the current or another speaker. If this initial context is not the optimal one for interpreting the upcoming utterance, the hearer must be pointed towards another context, or, at least, away from the expected, initial context. Two signals have been identified ('A framework for analysing smallword signals' pages 151 – 155) that point the hearer in a direction away from the initial context. The first involves signalling a break with the initial context when this is 'created' by the previous speaker – specifically the '*mode* changing' involved when a student embarks on a new task. The second signal in this section is that which indicates that the speaker's own immediately preceding utterance is no longer valid for interpreting what is to come; i.e. the context is broken in mid utterance.

Signalling a break with the initial context created by the previous speaker ('mode changing')

DEFINING EVIDENCE

What a new speaker says is normally interpretable in the light of what the previous speaker has just said. A smallword, typically *well*, may be employed as a signal when this interpretation is not entirely apt – i.e. some kind of break occurs between what has just been said and what is to come. This is most easily recognised in the data when a student embarks on doing a task, such as a narrative, when the immediately preceding exchange has been about the task. The student 'changes mode', switching from conversant to 'performer' and what s/he says is not directly interpretable in the light of the previous exchange. The contextual slot typical for signals of mode changing is that prefacing a new test task, as in [8] or a 'task within a task' (defined as giving directions or providing information or an explanation in a role-play task), as in [9]. *Well* occupies this slot in both examples:

[8] T: the pictures tell the story , you know , they start at the beginning , so just tell what happens , all right
A: **well** , this girl's going to go out , she's been asked to go out with her friends (43: 13 – 14) (NS)

[9] A: <reads> and can you describe the luggage please </>
B: **well** . it's two large cases one's black and one's green , and I've got my details on , on a tag (73: 84 – 85) (NS)

ANALYSING EVIDENCE

What emerged from 'Signalling an oblique response' pages 192 – 193, has already given an indication that the NoA students were aware of the potential of *well* to signal some kind of suspension or renegotiation of the initial context as the one most suitable for interpreting a response. This section examines students' use of *well* in situations where a more complete break is made with the initial context, which is created by what has just been said by the preceding speaker. It must be emphasised that the previous utterance alone cannot create a total context; however, the initial context being studied in this case is that which existed when the previous speaker finished speaking, as opposed to at some point in the present speaker's turn, which is considered in 'Signalling a mid-utterance break with context created by the speaker's own immediately preceding speech' pages 196 – 200.

Studies of the use of discourse markers in topic shifting have suggested that *well* may be a key player (e.g. Jucker (1993: 446) and Stenström (1994: 155) in signalling a shift in topic. However, neither of these studies touches on the role of a marker when the shift is not so much in topic as in something related to genre and speaker role, or 'mode' as it is called here. This occurs frequently in the data for this study because of the nature of the speaking-test situation. Students were thrown into certain tasks, and the points at which they ceased to be 'just talking' to the tester and became performers are the focus of interest here. Schiffrin (1987: 110) notes the use of *well* in introducing stories that respond to questions, but regards this use as signalling a deferral or a temporary suspension of context.

Table 8.3 shows the distribution, for the three student groups, of smallwords used at a *mode* change, prior to starting a task (stories, descriptions, instructions and explanations) or a sub-task in the role-play.

Table 8.3 Raw data on smallwords in the contextual slot for signalling 'mode changing' prior to a task

smallword	**NS**	**NoA**	**NoB**
right	14	-	-
all right	1	1	-
okay	1	16	15
well	24	7	5
oh	-	1	-
ah	-	1	-
total	**40**	**26**	**20**

As all students were in principle given an equal opportunity to perform the various tasks, it is most reasonable to consider the figures in the table relative to the numbers of students in the groups: NS=18, NoA=19,

NoB=24. While the NS students on average prefaced a task with a smallword more than twice each, this was only done on average rather more than once by each of the NoA students and rather less than once by each of the NoB students. In their choice of smallword prior to mode changing, the NS students almost exclusively used *well* and *right,* whose percentages of the overall smallwords in this context were roughly two-thirds and one-third respectively. Both Norwegian groups kept to *okay* (making up about three-quarters of the smallwords) and *well* (approximately one quarter).

A more detailed study of the NS data for individual task types (not indicated in Table 8.3 on page 195) suggests that the distribution of these smallwords was apparently not random. While *well* was favoured (making up three-quarters of the smallwords) to preface narratives, descriptions and explanations, *right* was favoured to preface instructions, where *well* made up only one-third of the smallwords used.

It must be concluded from these data, that the native-speaker students tended to signal mode changing with smallwords when embarking on narratives, descriptions, explanations and instructions. They did this on just over half of the potential occasions, on average. What is more, they definitely tended to use *well* to give this signal, although in the case of instructions their preference was for *right.*

The data for the Norwegian students, on the other hand, does not contain the evidence to suggest that they generally used smallwords to give this signal, although the NoA group used rather more smallwords per head than the NoB group to preface a mode change. They seemed only to use a smallword in this context on about a quarter of the potential occasions. *Right*, as has been observed in the case of other signals, was not represented, and *well* was only thinly used, with *okay* once again being the favourite choice.

Signalling a mid-utterance break with context created by the speaker's own immediately preceding speech

DEFINING EVIDENCE

Speakers sometimes break the flow of their own utterances and frequently signal this by means of a smallword placed at the break point, i.e. the point where they break into the original message and embark on the bit of 'new message' that does not quite follow from the original. This point is regarded as the typical contextual slot for signalling a mid-utterance break. Four kinds of mid-utterance break are recognised in the data. The first is when the speaker gives some sort of digression, aside or metacomment, as in [10], where *well* occurs at the break point:

[10] B: it's very safe it never happens anything here , **well** I don't know about <place name> , I think it's quite safe there too (50: 63) (NoA)

The second is when the speaker returns to the main topic after a digression, as in [11], with *well* again occurring at the break point:

[11] B: ... ceiling like big worms and in the end I couldn't hear the tunes 'cause my ears were ringing that loud . em , **well** me and my friend were both really excited afterwards , ... (44: 53) (NS)

The third kind of break occurs when an idea has apparently suddenly struck the speaker and is added into the text rather abruptly, clearly not following smoothly from what has gone before. This typically occurs in task 1 of the test, during the picture description, where a detail is suddenly noticed, as in [12], with *oh* at the break point:

[12] A: in a club , and a bloke comes over and he says . can I get you a drink . and it's a general sort , **oh** he's just come from a table that's empty now , and it's a general sort of (41: 6) (NS)

In [13] the student is struck by a thought after the other student has given a minimal response, *yes*, but it is regarded here as a continuation of the speaker's utterance, as he is still 'holding the floor'. Again, *oh* occurs at the break point:

[13] A: and then you know where his dog bowl is
B: yes
A: **oh** and you'll have to er open the food , you know where the tin opener is (74: 94 – 96) (NS)

The fourth type of break occurs before a 'self-repair'. Speakers monitor their own speech, and frequently realise that this needs to be repaired in some way. Levelt (1983: 51 – 53) identifies three types of self-repair, corresponding to the questions: *Do I want to say this now?*, *Do I want to say it this way?* and *Am I making an error?*. Of relevance to this study, Levelt observes that self-repairs are characterised by the use of 'editing terms', which, '[together with] the first word of the repair proper, almost always contain sufficient information for the listener to decide how the repair should be related to the original utterance' (1983: 41). Although Levelt does not cite smallwords in this capacity, these editing terms can presumably be

interpreted as having the signalling function being discussed here, and therefore to include smallwords.

Self-repairs in the data, corresponding to Levelt's three types, are recognisable as: a (partial) restart with a reordered message, as in [14], a (partial) restart with a different formulation, as in [15], and a correction, as in [16]. As can be seen, both *well* and *oh* are placed at the break points:

[14] T: ... just say what's happening
A: well , I think they are , these two are , they're going to see a match and , **well** , eh , they are now in the , in a station café and , eh , (44: 18) (NS)

[15] B: his son would do such a thing . I think that's what most , **well** the parents react if something like that happens (50: 53) (NoA)

[16] A: all right . well , they've gone to the café , and then they go to buy tickets . **oh** no , he asks her where the tickets are , and she says ... (49: 8) (NS)

ANALYSING EVIDENCE

The data shown in Table 8.4 illustrates the smallwords that occurred in the contextual slots for mid-utterance breaks of the types described above. While the native-speaker students provided 15 smallword tokens in this slot, the NoA students provided a rather lower 11, and the NoB students only six. It is clear from Table 8.4 that virtually all of the Norwegian occurrences took place before a self-repair. The native speakers produced fewer tokens in this particular context, understandably, perhaps, under the circumstances, where they were less likely than the learners to need to repair what they said.

Table 8.4 Raw data on smallwords occurring in the contextual slot for signalling a mid-utterance break

smallword	**digression/aside**			**sudden idea**			**self-repair**			**total**		
	NS	**NoA**	**NoB**	**NS**	**NoA**	**NoB**	**NS**	**NoA**	**NoB**	**NS**	**NoA**	**NoB**
well	3	1	-	-	-	-	1	6	2	**6**	**7**	**2**
oh	-	-	-	6	-	1	1	2	1	**7**	**2**	**2**
ah	-	-	-	-	-	-	1	1	-	**1**	**1**	**-**
I mean	-	-	-	-	-	-	-	-	1	**-**	**-**	**1**
I know	-	-	-	-	-	-	-	1	-	**-**	**1**	**-**
okay	-	-	-	-	-	-	1	-	1	**1**	**-**	**1**
total	**5**	**1**	**-**	**6**	**-**	**-**	**4**	**10**	**5**	***15***	***11***	***6***

The table shows that only *well* and *oh* recurred to any noticeable degree in this contextual slot. Since the NS group produced both smallwords more or less equally, it can be concluded that they used both of these smallwords to signal a mid-utterance break. The NoA group's greater tendency to use *well* may be explained by the fact that most of this group's tokens occurred in self-repair contexts. Most of the *oh*s supplied by the NS group occurred in 'sudden idea' contexts. It is therefore difficult to draw any conclusions on the nativelikeness of the NoA group's selections of *well* and *oh*. However, the evidence seems to suggest that the more fluent students behaved in a nativelike way, at least in their apparent use of *well* to signal a mid-utterance break.

The documented use of *well* by native speakers to signal a break with the context of the other speaker's utterance was discussed in 'Signalling a break with the initial context created by the previous speaker ("mode changing")' pages 194 – 196. That it should be used to give a similar signal during one's own speech is unsurprising. Schiffrin, who treats *well* as a marker of response, explains this phenomenon thus: 'speakers are treating their own prior talk as something to be responded to' (1987: 123). Svartvik (1980) (who also treats *well* as a response marker) comments on the high frequency of anacolutha occurring after *well*. Stenström (1984: 141), Jucker (1993: 446) and Stenström (1994: 85) comment on the role of *well* as a frame, which may indicate a partial or complete topic shift.

In other words, *well* is typically used during one's own speech when some sort of break in flow occurs, and the listener needs to be aware that what is coming cannot be directly interpreted in the light of what has just been said. Jucker sums up the use of *well* in both this and the previous section thus: '... *well* can be seen as a signpost signalling to the hearer that the context created by the previous utterance – whether produced by the current speaker or the current listener – is not the most relevant one for the interpretation of the impending utterance' (1993: 440).

The preference of the NS student group for using *oh* in a situation where the mid-utterance break is triggered by something 'external' to what is currently being said reflects the conclusions of Heritage (1984) and Schiffrin (1987), in their discussions of *oh* as a marker of change of cognitive state. Heritage states that evidence from conversational (native speaker) data shows that 'the particle is used to propose that its producer has undergone some kind of change in his or her locally current state of knowledge, information, orientation or awareness' (1984: 299). This use of *oh* signalling a change of state resulting from the other speaker's utterance is discussed in 'Signalling a cognitive change of state, resulting from the previous utterance' pages 201 – 204.

However, a change of information state, according to Schiffrin, may be 'cognitively triggered by the speaker's own processing of information, or

contextually triggered by an event' (1987: 95). An extension of Schiffrin's claim to include Heritage's range of 'states' that may be signalled as 'changed' yields an account of the use of *oh* which accommodates the examples in the student data being studied. Thus cognitive triggering is illustrated by [13], when student A remembers that he should tell his friend how to open the tin of dog food. Contextual triggering occurs in [12], where student A notices a detail in the picture she is describing. Aijmer (1987) sums up this use of *oh* as follows:

> *Oh* and *ah* can be associated with an interruption or intervention in the conversation at the point at which a person reacts to an unexpected situation. This is the case if the speaker suddenly has a certain insight but also if he guesses or infers something, or successfully solves a problem. Oh (ah) construes what comes afterwards as topically not coherent(...). It signals a shift or development to something not foreseen by the speaker. (1987: 63)

By using *oh* at these points in the conversation, the speaker thus signals that he has undergone some kind of change in what is at the forefront of his mind, and the hearer should therefore not expect to interpret what comes after the *oh* in the light of what was said before it.

Thus it seems that the NS student group used *well* and *oh* to signal a mid-utterance break in a way that reflects native-speaker usage generally. The more fluent Norwegian students showed that, in the case of self-repairs, they were able to signal this break in a nativelike way, using *well.* The weaker students did not do this, despite the fact that they were, presumably, at least equally in need of doing so. This result can be compared with the findings of Fulcher (1996), which reveal that self-repair, or reformulation, is a characteristic of more advanced speakers (bands 4 and 5 on his scale) (1996: 237 – 238). The fact that neither Norwegian group used smallwords to signal the other types of mid-utterance break suggests a gap in the signalling ability of both groups, since it is unlikely that they were significantly less in need of having to make such breaks than the native speakers were.

Indicating the cognitive effect of the previous utterance

In 'A framework for analysing smallword signals' pages 151 – 155, it was established that the default reaction to, or 'cognitive effect' of, what the previous speaker has said is that the hearer's assumptions are confirmed or added to in a somehow unsurprising, anticipated way. When a different, non-default, kind of cognitive effect has taken place, some signal might be expected, and two such signals are identified. The first signal indicates that a cognitive change has taken place in the now current speaker as a result of the previous utterance. The second signals that the inferences intended to be

communicated by the previous speaker have been rejected by the hearer. As was stated in 'A framework for analysing smallword signals' pages 151 – 155, no evidence has been found in the current dataset of the latter type of signal being sent. This is, doubtless, due to the nature of the testing, where students presumably did not feel entirely free to reject what was said out of hand. Because of this, the only type of signal considered here is the former type, signalling a positive change of mind-set, in line with that of the previous speaker.

Signalling a cognitive change of state, resulting from the previous utterance

DEFINING EVIDENCE

When the effect of what the previous speaker has just said is to add significantly to or replace what the current speaker assumed before, rather than to 'fit in' with existing assumptions, a signal of this may be given. *Oh* has been cited in the literature as signalling this 'change-of-state' (Heritage 1984) or 'shift in orientation to information' (Schiffrin 1987). This use of *oh* to signal this change of state when it occurs during a speaker's own turn was discussed in 'Signalling a mid-utterance break with context created by the speaker's own immediately preceding speech' pages 196 – 200. The signalling of such a change taking place as a result of what the previous speaker said is given considerable attention by Heritage (1984), who demonstrates how *oh* 'is unique in making a change-of-state proposal which is most commonly used to accept prior talk as informative' (1984: 335). This is illustrated in [17], where the tester breaks into a task routine with a piece of 'genuine' information about English phone answering routines, which the student was clearly not aware of previously:

[17] B: okay , but what is this
T: that's the phone number yeah it's quite normal in England
B: **oh** yeah (03: 299 – 300) (NoA)

However, there are many occasions where information, however 'new', cannot seriously be considered to cause a significant change in the 'mindset' of formerly held assumptions. Responses to questions such as *what's your name?* or a set of 'asked for' instructions are a case in point, and can be considered examples of what Schiffrin (1987: 89) calls 'anticipated' new information. Schiffrin uses this term with reference to the answers to questions, where the answer is selected from the 'question-encoded options'. Anticipated new information will be interpreted in the widest sense here, to include any answer that falls naturally within the range of those that can be

reasonably expected, as in [18]. It will also be extended to cases where the response is not prompted by a question, but where the situation demands that a certain type of information is expected, as in [19].

A smallword produced on receipt of 'new unanticipated information' is regarded as occurring in the contextual slot for signalling a cognitive change-of-state brought about by the receipt of the information. Information will be regarded as new if it can be reasonably assumed that the 'receiver' could not have known, or worked out for himself, what he has just heard. However, information is regarded as anticipated (i.e. it does not count as constituting evidence of this signal) in the following cases:

- it is judged to fall within the range of answers to a question that may be conventionally expected, as in [18]
- it is part of the normal, task instruction/assistance/explanations from the tester, as in [19]
- it is given by another student as an 'expected' part of a task, e.g. a set of instructions, as in [20].

Thus *right* in [18] and *okay* in [19] and [20] are not counted as occupying the contextual slot for this signal:

*[18] B: <reads> ((what size do you take </>))
A: I take size seven
B: **right** <reads> ((well here are all the sports shoes in that size . can you see any you like </>)) (41: 174 – 176) (NS)

*[19] A: yeah , first there is a boy and a girl I don't know there's name but
T: well they're called Steve and Ann , Steve and Ann
A: **okay** (01: 26 – 28) (NoA)

*[20] A: which button should I pull when I'm going to start the machine
B: there is ... there is a button you should decide which program and it's a lot of programs a b c d and so on and you should decide programme b
A: **okay** (15: 135 – 136) (NoA)

ANALYSING EVIDENCE

The raw data on smallwords occurring in the contextual slot for signalling a cognitive change of state, resulting from the previous utterance, is shown in Table 8.5.

The numbers of these occurrences are rather low. Yet they indicate that the NS and NoA speakers used smallwords in this slot to a roughly similar

Table 8.5 Raw data on smallwords occurring in the contextual slot for signalling a cognitive change-of-state, resulting from the previous utterance

smallword	NS	NoA	NoB
oh	5	10	5
ah	3	-	-
total	**8**	**10**	**5**

extent, relative to the total numbers of words uttered, with the NoB group using them relatively less. However, another feature arises from this sparse set of figures, i.e. that the NS group used both *ah* and *oh* when a cognitive change of state occurred. This particular phenomenon may be explained by the contextual situation that most often yielded the smallwords concerned here, namely a role-play telephone call from a stranger announcing that he is about to visit the student's home. This is a situation yielding 'unanticipated' information as defined above, while at the same time demanding a 'positive' reaction.

Aijmer (1987), in her study of *oh* and *ah*, reports that *ah* was used much less frequently than *oh* in her data from the London-Lund corpus (with 77 *ah*s compared with 716 *oh*s). However, her account of the use of *ah* may help to explain its use here. She observes that *ah* 'conveys in addition a sensation of pleasure when the speaker (suddenly) observes something he has been looking for' (1987: 65). Furthermore, she cites James (1978) who finds that the difference between *oh* and *ah* is that *ah* 'seems to always indicate that the speaker is pleased or that he thinks that the thing he has found out is significant in some way' (James 1978: 519 in Aijmer 1987: 66).

This 'flavour' added by *ah* to signal that information is not only new and unanticipated at the moment it comes, but also pleasing, significant and in some way 'looked for', is reflected in the way it is used in example [21]:

[21] B: <reads> hello , do you speak English please</> <156>
A: yes , a little
B: er <reads> well , my name's Stephen White , I'm on my way to visit you</>
A: **ah** yes er my father said you should be we're expecting you soon (74: 155 – 158) (NS)

The NS group used *ah* twice and never used *oh* in this particular situation. The NoA group, on the other hand, never used *ah* but used *oh* three times in this situation, as illustrated in example [22]:

[22] B: <reads> well , my name's Stephen White . I'm on my way to visit you</>)
A: **oh** , you are
B: <reads> ((could I speak to your father please</> (14: 207 – 209) (NoA)

Moreover, the NS students were inclined to use *ah* when expressing sympathy in general, such as in the role-play when the other student tells about his/her ordeal outside the cinema, as in [23]:

[23] A: <reads> ((you know I waited half an hour then went home , it was awful </>))
B: **ah** I'm really sorry , listen do you want to come round tonight to my house (42: 135 – 136) (NS)

In fact, before *sorry*, *ah* occurred four times (although only once when giving this change-of-state signal) in the NS data, with no occurrences of *oh*. This suggests that *ah* was used to send a signal of empathy. Only on occasions where there seemed to be no genuine or pragmatic reason (e.g. for politeness) to display pleasure, did the NS group mark the receipt of new information with *oh,* as in [24]:

[24] T: yeah , well you know , act as if you know you've had him on the phone and these are the things he needs to know
A: **oh** I see . so he's come by boat (72: 42 – 43) (NS)

Any conclusion to be drawn from this section can only be very tentative, with such sparse evidence. Yet, the indications are that both Norwegian groups showed awareness of the potential of *oh* in signalling a cognitive change of state as a result of the previous speaker's utterance. However, these students were apparently not aware that, by using *ah* instead of *oh*, they could have 'flavoured' this signal either to show that this change of state was a 'positive' one, or to express some sort of empathy. Thus, the nativelikeness of the Norwegians' signalling a cognitive change-of-state was restricted, with possible pragmatic consequences.

Indicating the degree of vagueness or commitment: Signalling a softening of the impact of the message, or 'hedging'

In this section, dealing with the macrosignal, which indicates the degree of vagueness or commitment in an utterance, only one signal is identified, that

of signalling a softening of the impact of the message, or 'hedging'. However, this single signal occupies a major place in this study. Not only does it involve eight of the 19 smallwords studied here, but it also accounts for around 550 of the total of approximately 1,200 tokens of smallwords registered. In other words, hedging accounts for almost half of the smallword use studied here. For this reason it is given fairly extensive and detailed coverage.

DEFINING EVIDENCE

The function of hedging (as it was defined in 'A framework for analysing smallword signals' pages 151 – 155) is regarded here as either expressing 'vagueness' within the proposition itself, e.g. by using *sort of*, or expressing a down-toning attitude or a lack of commitment to the truth of the proposition, e.g. by the use of *just* or *I think*.

Whereas in the case of most signals, evidence is defined purely by the number of occurrences of smallwords in specified contextual slots, the definition of hedges is based solely on the inherent meaning of the smallwords themselves. *Well* might be used to signal vagueness, but occurrences of this happening are not looked for here because of the difficulty of defining a typical context for the signalling of vagueness. The following smallwords, because of their transparency, are considered to explicitly and inherently function as hedges (i.e. they always make the message less forceful than if they were omitted), and therefore do not depend on contextual clues in order to be recognised as such:

I think	like	sort of/kind of	a bit	just
or something	not really	and everything/that/stuff/things		

It must be emphasised, however, that only when they satisfy the conditions of 'smallwordness' – principally insofar as they might be dropped without affecting the syntax or the essential 'meaning' of the message – are these expressions counted as giving this signal. Moreover, occurrences of *just* are excluded when performing as an adverbial, as in *he's <u>just</u> arrived* or an intensifier, as in *it's <u>just</u> unbelievable!*

Hedges can be prompted by two distinct motives, as is made apparent in the course of this section. Firstly, hedging is well documented as being frequently motivated by 'interpersonal' considerations, such as 'face-wants' and empathising (see e.g. Brown and Levinson 1987, Channell 1994, Stenström 1994, Nikula 1996). Such pragmatic, interpersonal motivation probably prompted the use of *just* and *or something* in [25], where a student is asking a friend to do him a favour by exercising his puppy (and is presumably trying to follow the maxim 'don't coerce' (Brown and Levinson 1987: 172)):

[25] B: ... in the garden so it can run and take some fresh air , and if you're going to , walk er round with it you need er you can **just** follow it around , in some ten minutes on the street **or something** (03: 102) (NoA)

Secondly, hedging can be employed to soften a message for more 'genuine' semantic reasons, concerned with the way the speaker relates to the message itself, which can give rise to the need to disclaim any certainty or precision in what he is saying, as in [26]:

[26] A: er . his name is , <name> and he is fifteen years old . **I think** <laughs> (11: 55) (NoB)

Holmes (1984) sums up this duality in the motivation for hedging, stating:

> There are at least two basic reasons why a speaker may wish to modify the strength or force with which a particular speech act is expressed: firstly to convey modal meaning or the speaker's attitude to the content of the proposition, and, secondly, to express affective meaning or the speaker's attitude to the addressee in the context of an utterance. (Holmes 1984: 348)

No claim is made here to be able to distinguish between hedges which are motivated by proposition-related or affective, interpersonal considerations, not least because it is normally impossible to be sure of the motivation without access to the speaker's mind. However, because the ability to use hedges equips a speaker not only to affect the propositional content of what s/he is saying, but also to convey certain interpersonal signals, the absence of these items from her/his vocabulary can be assumed potentially to detract not only from the fluency, as it is under discussion here, but also from the pragmatic language ability of the speaker.

ANALYSING EVIDENCE

As was commented on in 'Expressing the communicative intention' pages 188 – 193, hedges made up about 80 per cent of the smallwords used in turn-internal position for all student groups, which bears out the contention made by Channell (1994): 'Most speakers of English are not aware of the frequency of vague language use (until it is pointed out to them) and this fact is of itself of interest. It shows that vagueness in communication is part of our taken for granted world, and that normally we do not notice it unless it appears inappropriate' (1994: 4).

The figures for smallwords used in turn-internal position (discussed in 'Range and variety in smallword use' pages 173 – 176) indicated that not only did the NS students use many more hedges than the Norwegian groups (180 per 10,000 words, compared to 80 (NoA) and 54 (NoB)), but the range of their hedges was also far more varied. The NS group used seven different

hedges regularly (i.e. more than five times), while both NoA and NoB groups used only four, with three (*just, I think* and *or something*) dominating, causing one to wonder how far these students' use of vague language can have been 'appropriate' in Channell's terms, with such a reduced repertoire to draw on.

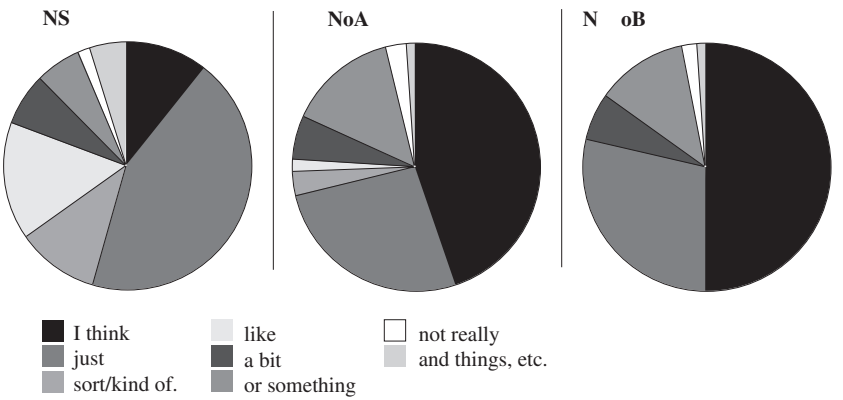
The raw data for smallwords used in signalling a softening of the impact of the message, or 'hedging', is shown in Table 8.6.

Table 8.6. Raw data for smallwords used in signalling a softening of the impact of the message, or 'hedging'

smallword	NS	NoA	NoB
I think	29	71	56
just	102	42	31
sort/kind of	24	5	-
like	46	3	1
a bit	17	9	6
or something	17	24	14
not really	4	5	2
and things...	14	2	2
total	**253**	**161**	**112**

The number of hedges used shows clearly that while neither Norwegian group 'matched' the NS students in their use of smallwords generally to give this hedging signal, both Norwegian groups used smallwords to give this signal, with the NoA group outnumbering the NoB group, although not on the scales that the raw data indicate, as the NoA group produced many more words. A more balanced view of the way hedges are distributed among the three groups is indicated by Figure 8.2, which shows the proportions of smallwords used by the three student groups to signal hedging.

Figure 8.2 Proportions of smallwords used by the three student groups to signal hedging



The most striking difference between the native-speaker chart and those for the Norwegians is that the NS circle is more evenly divided up than the Norwegians', where about three-quarters of the 'pies' are made up of two smallwords only, viz. *I think* and *just* (which, admittedly, make up half of the NS pie).

The narrowness in range of hedging smallwords used by the Norwegian students, while not essentially reducing the overall message of 'vagueness' or 'softening' in a message, can have two principal effects on the students' ability to communicate smoothly. The first is simply that the language may be perceived as monotonous. The second, more subtle, effect may be that certain additional functions performed by individual hedges may be put out of the reach of students with limited ranges of hedges. The following sections will hopefully shed some light on the second of these effects.

Learner-favoured hedges

Although the number of tokens of hedges was generally fewer among the Norwegians, these students tended to cling to certain smallwords to do their hedging, sometimes producing many more tokens of these than the NS group did. As can be seen from Figure 8.2 on page 207, *I think* emerges as the overwhelming favourite hedge in general use among the Norwegians, while occupying a far less significant place among the NS choices. The general preference for *I think* among these students might be put down to the fact that it is a highly explicit way of expressing lack of commitment to anything from the choice of a single word to a whole proposition; the Norwegian students, like Nikula's non-native speaker subjects, opted for explicitness when selecting a hedge. It seems also to be perceived by these students as syntactically highly mobile, with students in both Norwegian groups sometimes adding it as an attachment, as in [27]:

[27] A: .... her dad is coming up and er , the man or the thief is getting shook **I think** (65: 13) (NoB)

The relatively great use of *or something* (the tokens of which also include *or something like that*) may have a similar explanation. Like *I think*, *or something* is transparent, as well as being capable of relating to the whole proposition. The majority of the Norwegian groups' tokens of this smallword occurred in turn-final position, in most cases seeming to be 'attached to' a whole proposition, rather than to a single element. In these cases one gets the impression the students used it to 'cover themselves' for some or all of what they had said, as in [28]:

[28] A: so example <unclear> she just tells me they are parking their car **or something** (17: 14) (NoB)

Another smallword the Norwegian students favoured as a hedge was *just,* although in this case the NS students used even more tokens. Lee (1987), in his analysis of the different uses of *just,* claims:

> the first and most obvious type of meaning expressed by *just* is one that I will call the deprecatory meaning. In examples illustrating this meaning the speaker uses the particle to minimise the significance of some process. (I use the term process as a cover term for events, actions and situations.) (1987: 378)

All three student groups produced many tokens of *just* used in this way, as in example [29], where B is talking about what he did at the weekends:

[29] B: ... um I , took my bike and went **just** around for a while , and did nothing I think I did my homework I think that's all (50: 65) (NoB)

Thus it appears that the apparent ability of both Norwegian groups to use *I think, or something* and *just* to signal hedging, equipped them to soften the force of their opinions, to cover themselves when they were unsure whether what they were saying or had just said was 'right', and to play down the significance of what they were talking about. However, by largely restricting themselves to these three smallwords, the Norwegians, particularly the NoB group, did so in a rather monotonous, and, at times, awkward way.

Learner-underused hedges

Smallword hedges used significantly by the native speakers, yet hardly at all by the Norwegians were *and things/everything/stuff/that* and *like*. Smallwords relatively 'underused' by one or both Norwegian groups were *sort/kind of, a bit* and *not really*.

Together with *or something*, which has already been discussed, the category of smallwords consisting here of *and things*, *and everything, and that* and *and stuff* is discussed under various names in the literature. Overstreet and Yule (1997), for example, label them 'general extenders', while De Cock *et al.* (1997: 76) use the term 'vagueness tags' in their comparison of native- and non-native-speaker use of these expressions. In the latter study, the non-native speakers were found to make very little use of these tags, compared to the native speakers. In the present study, a similar result is recorded, with only two non-native occurrences of the *and things,* etc tags, compared with 15 from the native speakers.

Overstreet and Yule (1997) ascribe functions on what can be regarded as two levels – the propositional and the interpersonal – to 'general extenders'. They point out that by attaching an extender such as *and things* to an item, a speaker can extend the meaning of the item to include any category it may belong to. This may be a conventionally identifiable or 'lexicalised' category,

or an 'ad hoc' one. The former type of category can be illustrated by example [30], where the use of *and that* extends 'football scarves' to cover all 'supporter gear':

[30] A: station and there's two people with all the football scarves **and that** on , and there's a taxi , taxis , two taxis , and there's a bus with the united supporters on , (45: 4) (NS)

An ad hoc category, 'the things you have to do when you come home with a dog', is implied by the use of *and everything* in example [31]:

[31] B: him for a walk , maybe in the park or somewhere , and you come home again , and you em , you wipe your feet **and everything** (45: 64) (NS)

Overstreet and Yule (1997) go on to maintain that the use of these general extenders may be inspired by the desire to empathise, or 'show solidarity', with the other speaker, indicating that we are making assumptions about shared knowledge and experience. Thus, in lacking this particular tag, not only were the Norwegians deprived of the opportunity to refer to a 'fuzzy' idea with little effort, but they might also have been less able to perform the important interpersonal function of empathising.

The other hedging smallword used virtually exclusively by the NS speakers was *like.* As in the case of other smallwords, such as *I think, like* is 'accepted' as a smallword in the data only when its hypothetical removal would not produce a semantically or syntactically distorted utterance. Thus in example [32] *like* is not regarded as a smallword, while in example [33] it is regarded as such:

*[32] A: um , it looks **like** it's in the railway station ... (45: 5) (NS)

[33] B: ... em , well me and my friend were both really excited afterwards , and **like** we formed a great fan club (44: 53) (NS)

Andersen (1998) illustrates that *like* in teenage native-speaker speech 'is viewed as a general marker of loose use of language which explicitly signals that the utterance in some respects is a less-than-literal rendering of a speaker's thought' (1998: 157). He argues that *like* encodes 'procedural information', signalling something about the procedure of how a listener is to interpret an utterance, and contrasts it with *sort of,* which normally denotes an actual concept.

The numerous examples in the NS data in this study lend support to Andersen's claim that *like* is widely used to signal that loose interpretation is

expected. *Like* in example [33] above seems to encompass the whole proposition *we formed a great fan club* as its object and could be taken to mean that the listener must not take this proposition too literally, in the sense of all that is implied in fan club formation. In example [34] below, *like* is also referring to a proposition, that people were taking a taxi, but here the looseness of the utterance lies in the fact that this is simply an example of the kind of things people were doing. In example [35], *like* has a more restricted scope, where the listener is being giving the signal that the things being referred to are only 'approximately' cupboards.

[34] A: and there are people , it's very crowded , there's so many people , some of them , **like** , have , taking a taxi or they're , they are , you can't see them in here but they are just standing (44: 16) (NS)

[35] A: ... and you can get off and put your luggage in ... , in these **like** cupboard things ... (45: 54) (NS)

This versatility of *like* makes it a most useful tool in signalling a 'looseness', or softening, of what we are saying. It can encapsulate our attitude to or belief in what we are saying, or the degree of precision in the way we are expressing ourselves, by simply signalling 'don't take this too literally'. Not surprisingly, it is a well-used, and arguably overused, tool in the hands of those to whom it is accessible. Schourup (1985), in his study of the use of *like*, accepts that its non-standard use is 'maligned' but states in its defence:

> In the present discussion it is argued that the current efflorescence of <u>like</u> in conversation, at least among younger speakers, is not a symptom, as Newman would have it, of the 'death of English', but the spread from its originally quite restricted range of occurrences of an item which in general indicates a possible loose fit between overt expression and intended meaning. (1985: 61)

While few might see it as desirable that Norwegian students should become '*like*-users' to the extent that some of the native speakers apparently were, there is little doubt that the Norwegian students, of whom only four out of the 42 used it at all, were deprived by apparently not being aware of its availability and potential.

In the case of *sort/kind of*, the five occurrences in the NoA group data illustrate that there was some awareness among these students of this smallword. However, considering that 24 tokens are produced by the NS group, the stronger Norwegians must be regarded as having underused this smallword, while their weaker counterparts did not use it at all.

Holmes (1988), in describing the use of *sort of* in expressing epistemic meaning, claims that it 'warns of impending imprecision' (1988: 116) in what is being said. This signal may be sent to show either that the concept being

expressed is to be regarded as an approximation, or that, for some reason, the speaker's choice of wording is not the most apt. Both of these types of signal are found in the students' data, as in example [36], showing the former type, and [37], illustrating the latter:

[36] T: so , where would the first one fit into her story
B: **sort of , sort of** at the beginning where they're looking for the trains , and what train to get to go to the football match (45: 13 – 14) (NS)

[37] T: there is something
B: mm , **sort of** a chain
T: yeah that's sort of a chain yeah or a lead mm (01: 180 – 182) (NoA)

It can be concluded that *sort of* was widely used by the native-speaker students to signal imprecision in what they were saying. While there was some evidence of awareness of the potential of *sort/kind of* in giving this signal among the NoA students, this evidence is very thin, and it is totally lacking in the NoB data.

The situation regarding *a bit* is perhaps more encouraging. While only 15 tokens were recorded for the NS students, the NoA group provided nine and the NoB group six tokens, showing some kind of comparability with the native speakers. Interestingly, *a bit* was produced in five of the six NS shoe-shop role-plays, where a student is telling the assistant that the shoes being tried on are too small, as in example [38]:

[38] B: er yeah sure , you can just try them on now
A: er , these are **a bit** small can I try a larger pair (71: 73 – 74) (NS)

Only one NoA token was produced in this situation, and not a single NoB token. Example [39] shows the typical Norwegian response without *a bit:*

[39] B53: yeah, just try them on
A53: um , they are too small, um can you find a bigger uh pair (53: 131 – 132 (NoA)

This rather bald statement, *they are too small,* exemplified in [39], in fact occurred eight times in the Norwegian data. This suggests that *a bit*, seemingly perceived as a necessary face-saving marker in this situation by the native speakers, was not perceived this way among the Norwegians.

The number of tokens of negator+*really* and *oh* as hedges was too small to warrant any conclusions about their relative use among the groups.

Conclusions on hedging

Several conclusions can be drawn from the evidence uncovered in this section. The most obvious conclusion is that, while all the groups used smallwords to signal hedging, the NS group did so in far greater numbers and using a much wider range of smallwords than the NoA group, who in numbers at least, if not in range, outperformed the NoB group.

Certain smallwords were heavily relied on in hedging by both Norwegian groups, notably the explicit and syntactically mobile *I think* and *or something*. On the other hand, some smallwords, particularly (the admittedly stigmatised) *like* and the *and everything* family of 'general extenders', were virtually missing from the speech of both Norwegian groups. Others, like *sort of* and *a bit,* were used to a lesser degree. By not using, or underusing these hedges, specialist functions were lost to these students, which deprived them of tools with particular potential for foreign language speakers. These include *sort of,* which can indicate that the choice of wording is not entirely apt, *and things,* which can extend an example into a category and *like*, which can tell the listener to interpret a whole utterance loosely. Moreover, certain pragmatic functions were denied them, such as the empathising implied by general extenders, and the face-saving marking of *a bit*. The less frequent use and the restricted range of their hedges are thus likely to have had an effect on not only the fluency but also the pragmatic ability of these students.

Indicating the state of success of communication

Two signals within the macrosignal of making reference to the state of success of communication have been defined: 'acknowledgement' and 'appeal'. By sending the first signal we let the other speaker know that we are getting the point of what s/he is saying. The second signal appeals to the interlocutor either to confirm that s/he understands what we are saying, or to help us with a difficulty in expressing what we mean. Defining typical contexts for these signals has not been straightforward, involving the inherent meaning of smallwords, and even intonation at times. Table 8.7 on page 214 shows the raw data on smallwords satisfying the various conditions for potentially signalling something about the state of success of communication.

Signalling the acknowledgement of smooth communication

DEFINING EVIDENCE

We can acknowledge that the speaker's message is 'getting through' without our actually taking the turn. This is done through 'backchannels' (like *m, okay, I see*) which do not involve speaker shift (see Stenström 1994: 5). Speaker T remains the speaker, as when *okay* is uttered in [40]:

Table 8.7 Raw data on smallwords used to signal the state of success of communication

	acknowledgement			appeal		
smallword	**NS**	**NoA**	**NoB**	**NS**	**NoA**	**NoB**
right	39	1	-	1	1	-
all right	4	4	5	1	-	-
okay	21	72	25	2	2	1
well	-	-	-	-	2	-
oh	-	3	9	-	-	-
ah	4	1	1	-	-	-
I mean	-	-	-	-	-	-
you see	-	-	-	4	1	-
you know	-	-	-	4	2	3
I see	-	2	1	-	-	-
I know	1	2	-	-	-	-
total	**69**	**85**	**41**	**12**	**8**	**4**

[40] T: to follow but remember he he doesn't know your house at all so you've got to tell him where everything is))
B: **okay**
T: and and if you need to if you feel that it's not quite clear you can just ask okay (01: 159 – 161) (NoA)

However, we may acknowledge what has been said by taking the turn (briefly) before passing it to a different speaker, as when *right* is produced in [41]:

[41] A: ... and then when you're finished you can go to the station and catch the train , which'll take you to Birmingham
B: **right**
T: do you want to ask him anything about this , ... (72: 52 – 39) (NS)

Smallwords produced in loner position, i.e. accompanied only by another smallword or *yes/no*, will be regarded as acknowledging the smooth flow of communication, provided that their inherent meaning seems conducive to this. In other words, all occurrences of *right, all right, okay, I see* and *I know* as loners are included. Occurrences of *oh* and *ah* are counted, provided that there is no evidence that the message does not seem to be getting across, as in [42] where it is rephrased by the first speaker:

*[42] T: and you could perhaps think of er , Oslo Centre [you know]))
B: [**oh**]
T: you know not just round here but but Oslo as a place to be , er (19: 87 – 89) (NoA)

Occurrences of *well* are not counted. As Stenström (1994) puts it, *well* signals 'hesitation, or doubt, or scepticism and so on' (1994: 113). It is clearly interpreted by T as expressing this doubt in [43]:

*[43] T: can you , did you catch that do you remember a bit of that
B: **well** yes
T: could you , would you like to just check <laughs>)) (71: 44 – 46) (NS)

Additionally, smallwords in non-loner position may signal the acknowledgement of smooth communication, typically following either an appealer, such as */all right* with rising intonation, given by the other speaker to check that communication is going smoothly, or an explanation that is explicitly given to assist communication. These two positions are therefore also defined as contextual slots for signals of acknowledgement, illustrated by examples [44] and [45], where they are both occupied by *okay*:

[44] T: ..... if he doesn't understand you he asks you if you ... if you're looking for a word he can help you **/all right**
A: **okay** eh okay you take the boat from Oslo to Newcastle and when you are at Newcastle ... (A14 166) (NoB)

[45] T: well that it's called a lead
A: a lead **okay** (02: 239 – 240) (NoA)

ANALYSING EVIDENCE

The data in Table 8.7 suggests that the signal most frequently sent by smallwords when referring to the state of success of communication was that of acknowledgement. Moreover, the absolute majority (between 74 and 80 per cent) of tokens occurring in the textual slot for acknowledging from all three student groups were loners.

It was established in 'General smallword use: quantity and distribution' pages 170 – 173, that there was no significant difference between the number of loners relative to total words supplied by the NoA and NS student groups, while the NoB students produced relatively few. The figures in Table 8.7 confirm that these facts apply equally to acknowledgers generally, as well as the fact that, once again, while the NS students preferred to use *right* in this function, only using *okay* on less than one-third of occasions, the Norwegian students overwhelmingly favoured *okay,* and virtually never used *right*. *All right* makes up a handful of acknowledgers from all three groups.

Interestingly, Stenström (1990), in her study of 'the *right* set', found that *okay* was used almost as often as *right* in the position of 'separate turn' in the

conversation of British adults in the London-Lund Corpus. There can therefore be no grounds for condemning the Norwegian students as 'wrong' in their favouring *okay* in this position. However, the fact remains that *right*, the very word most frequently used by the native-speaker students to give this signal, was missing from the vocabulary of the Norwegian students.

While tokens were very low for *ah,* it is perhaps worth commenting on the fact that, once again, the NS students selected *ah* rather than *oh,* which was favoured by the Norwegians. This preference has already been commented on in 'Signalling a cognitive change of state, resulting from the previous utterance' pages 201 – 204. There it was maintained, citing Aijmer (1987), that the use of *ah* signals that the receipt of a piece of information is pleasing or significant. Its use in signalling the acknowledgement of assistance in 'smoothing out' communication problems is illustrated in [46]:

[46] A: the where
B: the place where you can eat , have something
A: **ah** yeah , right (72: 82 – 84) (NS)

Aijmer (1987) offers a possible explanation for this use of *ah,* based on her study of *oh* and *ah* in the London-Lund Corpus:

> *Oh* expresses a reaction to an unexpected piece of information the significance of which need not at the moment be apparent to the speaker. *Ah* on the other hand implies the speaker now (at last) sees the significance of something that has eluded him before. (1987: 65)

By using *ah,* speaker A in [46] may somehow contribute to the smoothness of communication by suggesting that the explanation is in line with what the speaker might have expected.

The conclusion from the evidence in this section seems to be that, in terms of the quantity of smallwords used to acknowledge smoothness in communication, the NoA students performed on a par with the native speakers, while the NoB students lagged behind to some extent. However, when it came to their choice of smallwords, both Norwegian groups once again revealed an over-dependence on *okay*, at the expense of *right*, as well as a lack of awareness of the additional potential of *ah* in marking 'positive' overtones in the acknowledgement of what has been said.

Signalling an appeal to the listener to confirm or assist smooth communication

DEFINING EVIDENCE

Typical ways of appealing to the interlocutor to assist us in saying what we want, or in confirming that we are 'getting through' have not proved capable

of definition in terms of position alone. Intonation is also involved, and, to some extent, the inherent meaning of the smallword. Two distinct ways of signalling an appeal to the listener to confirm or assist smooth communication are recognised here. The first is usually performable by a smallword on its own, such as */all right* with rising intonation. The speaker uses this means to check that the hearer is 'getting the message', or that what he himself is saying is somehow acceptable. Typically, this is recognised by the rising intonation and the fact that some kind of response from the other speaker follows immediately, even being perceived as overlapping by the transcribers. Both of these conditions are satisfied in examples [47] and [48], by */okay* and */right*.

[47] A: she don't know my where my room [**/okay**]
T: [no] no she doesn't know where that is she doesn't ... (16: 125 – 126) (NoA)

[48] A: right . and so , you put your dirty clothes inside it **./right** , [and then] you go and you get the tin , which is next to the washing machine
B: [yeah]
A: which is got , which you'll see Ariel (42: 75) (NS)

The second way of using appealers occurs when a speaker is having difficulty in expressing what s/he means or when s/he, for pragmatic reasons, e.g. to empathise, prefers to let the other speaker work out what s/he means. This signal is typically given by *you know* and *you see*, where the smallword is regarded here as inherently sending this appeal. Frequently the response it elicits also suggests that the hearer has been successfully involved in working out what he means, as in [49]:

[49] T: you wash clothes
B: no wash, **you know**
T: cleaning yeah (57: 71)

ANALYSING EVIDENCE

The figures for smallwords that satisfy the conditions for signalling 'appeal' are low, and thus do not warrant any confident conclusions. However, the total numbers of these appealers relative to the total word counts seem to suggest that both Norwegian groups gave fewer such signals than the NS group. Concerning which smallwords were used to give this signal, the thin spread of smallwords selected by all three groups makes any pattern-spotting very difficult. The only comment worth making is probably that the Norwegian groups hardly ever used *you see* as an appealer, while this was

'joint favourite', although admittedly with only four tokens, among the NS students. This may suggest that the Norwegian students were unaware of this particular appealer. Although further research would be needed to investigate this, the evidence points to an underuse by both Norwegian groups of smallwords meant to appeal to their interlocutors in assisting or confirming smooth communication.

Summary

The main task in this chapter has been to present and evaluate evidence of the Norwegian student groups' native-speakerlike use of smallwords to send certain signals. The signals investigated were shown in Chapter 6, to make a fundamental contribution to facilitating verbal communication according to the relevance theory account.

The amount of evidence that the two Norwegian groups used smallwords generally to send signals has been assessed, and compared with evidence derived from the NS group. Moreover, the selection of specific smallwords used to send particular signals by the Norwegian groups has been compared with those typically selected by native speakers, as revealed not only by the NS students' data but also by a selection of the literature on research into NS speech. The findings are summarised in Table 8.8, which gives an overview of the extent of the evidence found to support the NS-like use of smallwords by the NoA and NoB groups in sending signals.

A quantity of evidence for general smallword use in the Norwegian data is described as 'comparable with NS' when the number of tokens, relative to total words, was roughly similar to that of the NS group. As a rule of thumb, when this relative number dropped to between roughly one- and two-thirds of that for the NS group, the evidence is described as 'less than NS'. Below this level, evidence is regarded as non-existent. The evidence for specific smallword use is described as 'limited' when the range of smallwords typically used by native speakers was not fully utilised. It is regarded as 'very limited' when there is such heavy dependence on one or two smallwords that the NS range was hardly reflected at all. When the group has been found not to have been giving the signal or to have been giving it using smallwords downright 'differently', the evidence is regarded as non-existent.

Unlike the analyses carried out of the extent and distribution of smallword use, in Chapter 7, the findings in this investigation are not based on statistical significance. This has been deliberate and the reasons for it are complex. The present investigation has been of a consciously more subjective nature than the previous one, particularly in terms of how the typical context and other characteristics of a signal are defined, and how the quantity of evidence is interpreted. The point has not been simply to compare 'quantities' of evidence

Table 8.8 Summary of evidence of native-speakerlike use of smallwords, generally and specifically, to send signals

signal	evidence of NS-like use of **smallwords generally**		evidence of NS-like use of **specific smallwords**	
	NoA	**NoB**	**NoA**	**NoB**
turn-taking/holding	less than NS	less than NS	limited	very limited
oblique response	comparable with NS	-	comparable with NS	-
mode changing	less than NS	less than NS	limited	-
mid-utterance break	less than NS (mainly self-repair)	-	limited	-
cognitive change-of state	comparable with NS	-	limited	-
hedging	less than NS	less than NS	limited	very limited
acknowledgement	comparable with NS	less than NS	limited	limited
appeal	less than NS	less than NS	limited	limited

numerically, but rather to see if evidence exists at all that a signal is being sent in a nativelike way. It can in fact be assumed that the quantity of hard evidence, in many cases, under-represents the actual occurrences of signals, since some have undoubtedly slipped the net, owing to the rigorous definition of contextual slots. But although the numbers of tokens are sometimes so low that they would not warrant chi-testing, they are capable of showing quite clear tendencies, and the patterning of these tendencies is, in fact, supported by common sense. This study may thus be regarded as an indicator of a situation that begs further investigation at a deeper level.

As Table 8.8 shows, there is evidence that the NoA students used smallwords generally to a degree comparable with that of the NS students to send three of the eight signals investigated: *oblique responses, cognitive change of state* and *acknowledgement of smooth communication.* Of these, only *oblique responses* were signalled in an entirely nativelike way, by the use of *well.* The remaining five signals are all found to have been sent through smallwords by the NoA group, but to a lesser extent than by the NS students, and again, through the use of a relatively limited range of smallwords.

The NoB group did not seem to send any signals to a nativelike degree,

and only five of the signals seemed to be sent at all, with no evidence of the remaining three: *oblique response, mid-utterance breaks* and *cognitive change of state.* The range of smallwords selected by this group was always limited, usually to a favourite 'teddy bear' smallword.

Taking each signal in turn, a rough summary can be given of the effects of the limitations of the two Norwegian groups. While both groups showed themselves to be capable of signalling to the other speaker that they wished to embark on a turn or not let it go, they seemed to do this less spontaneously, and in a more monotonous way than the native speakers. The lack of variety in the way they did this might be a hindrance to combining smallwords, which native speakers frequently did, particularly when getting started. Moreover, this lack of choice also narrowed the options of simultaneous functions that smallwords might perform on other planes, e.g. expressing ideas or performing 'social' acts, besides defining turn-roles.

Because the more fluent speakers apparently had some competence in using *well,* they were able, when taking the turn, to warn the hearer that they were not going to give the response that might have been expected. Moreover, they were sometimes able to signal the degree to which what they said followed from the last utterance, whether their own or the previous speaker's (although they did not normally use *well* in mode changing). The less fluent group, in not actively using *well,* were apparently not capable of giving these signals.

Both Norwegian groups appeared, on the rather thin evidence available, to use *oh* to show that they had just heard something they didn't already know or anticipate. However, neither group was able to show, through the use of *ah,* that this 'change of state' was a positive one, or that they sympathised with the other speaker. This can mean that their responses could have been interpreted as indifferent or unsympathetic.

It appears that both groups were aware of the need to hedge, either for interpersonal reasons or to show that what they were saying was not exact. They also had a feeling for down-toning the importance of what they were saying. However, they had, at best, a limited range of hedges with which to signal these things, depending heavily on *I think*, *or something* and *just*. A result of this is, as far as fluency is concerned, that they lacked essential devices for making strategic use of their limited vocabularies. By lacking *sort of*, both groups lacked a way of signalling that a word choice was not the most apt. By lacking general extenders, such as *and everything*, they were not able to use an example to convey a whole category. And by lacking the much-maligned *like,* they lacked a versatile way of doing all these things. Moreover, many of these smallwords carry pragmatic overtones, which were also lost to these students. The acquisition of a marker such as *a bit* could make all the difference to both groups, cushioning statements that could otherwise sound like complaints.

The more fluent Norwegians showed themselves well able to acknowledge that they were getting the message of the other speaker, even though again their choice of signals was limited, mainly to *okay*. This gave them a considerable advantage over the less fluent students when it came to keeping the conversation running across speakers. On the other hand, the rather thin evidence suggests that they might not have been so competent at checking that the other speaker was getting their message.

The overall conclusion must be that the pattern NS>NoA>NoB, already established in Chapter 7, regarding the extent and distribution of smallwords, has been found to have been maintained in this investigation into the manner in which smallwords are used. Not only did the NoA group show more nativelikeness than the NoB group in the extent to which they used smallwords at all to send signals, but they also selected particular smallwords in a rather more nativelike way. Thus the hypothesis proposed in 'Data, hypotheses and research questions' page 185, that more fluent students would be found to have used smallwords in a more nativelike manner than less fluent students, appears to be corroborated by the evidence unearthed and presented in this chapter.

However, it must be noted that even the more fluent, NoA, students only totally matched the NS students in the sending of one signal – *oblique responses.* In most of the others they were limited in the range of smallwords they used, and for half of the signals investigated, the evidence of their sending a smallword signal at all was considerably smaller than for the NS group. This could mean, in practice, that a sizeable number of students in the more fluent group were not aware of the potential of smallwords for sending these signals. This is taken as corroboration of the second hypothesis, that even students regarded as more fluent had gaps and limitations in their signalling through the use of smallwords.

At the beginning of this chapter, it was stated that three pieces of information were to be sought in the analyses. The first, concerning whether the Norwegian groups used their smallwords in a reasonably nativelike way and the second, concerning the broader effects of the Norwegian students' smallword underuse on their fluency in performance, have been provided. The third piece of information concerned the way learners appear to acquire smallwords in their various signalling functions. It seems reasonable to assume that the NoA students were at a more advanced stage of acquisition of smallwords use than the NoB students. This theory is backed up by the fact that all smallwords used regularly by the NoA students were also used by the NoB students, while the reverse was not the case. Therefore, it is possible to put forward certain tentative proposals, based on what has been summarised above, on the way this use is acquired.

Table 8.9 shows three hypothetical stages at which the signalling functions of smallwords can be considered to be acquired. Stage 1 represents the stage

reached by a 'typical' NoB student. Stage 2 is that reached by a 'typical' NoA student. Stage 3 is that (ideally) reached by a student with nativelike fluency. A smallword's being acquired is loosely equated with its being used regularly, i.e. (at least) in quantities comparable with those of the native-speaker students. Smallwords are entered only in columns corresponding to the stage at which they are first used regularly.

Table 8.9 Hypothetical stages at which the signalling functions of smallwords appear to be acquired

	stage 1 (as exemplified by NoB group)	**stage 2** (as exemplified by NoA group)	**stage 3** (as exemplified by NS group)
turn-taking	*okay, oh, I think*	*well*	*right, ah, all right*
turn-holding	*just, I think, okay, or something*	-	*right* (and a variety of hedges: see below)
oblique response	-	*well*	
mode changing	*okay*	-	*well, right*
mid-utterance break	-	*well*	*oh*
cognitive change of state	*oh*	-	*ah* (with positive overtone)
hedging	*just, I think, or something,* (minus politeness overtones)	-	*sort/kind of, like, and things/that/everything/ stuff a bit*
acknowledgement	*okay*	-	*right, ah*
appeal	*you know*	-	*you see*

The order of acquisition of smallwords can be summed up as follows: *okay, just, I think, or something, you know* and most uses of *oh* are acquired early, by stage 1. Most uses of *well* are acquired by stage 2. Eventually, at stage 3, *right, all right, ah, you see* and the remaining uses of *well* are (all being well) acquired, in addition to the hedges: *sort/kind of, like, and things/that/everything/stuff* and *a bit*.

This chapter has thrown light on some broad tendencies in the way students of different fluency levels used smallwords to send signals essential to fluent speech. It has corroborated the theory that using smallwords in a more nativelike manner appears to go hand in hand with greater fluency. It has also revealed something about the way a reduction in smallword signalling affects the language, even of the more fluent group of speakers. Finally, it has yielded some information about the probable order in which smallword use is

acquired. This information is of significance for the enhancement of test descriptors of fluency. Moreover, the corpus analysis has yielded some detailed information about the order in which the learners studied appeared to acquire smallwords. The next chapter will draw on and add to the findings on smallwords cited here, with the ultimate aim of investigating their implications, not only for assessment but also for the teaching and learning of the spoken language.

9 The smallword user

This study has been preoccupied with smallwords as they appeared in a composite corpus of spoken student language. They have been counted and assigned signals, and the speakers themselves have only been of interest insofar as their apparent level of fluency was concerned. In this chapter some questions will be raised, and briefly addressed, that bring the speaker – i.e. the smallword user – into focus. Three issues will be taken up: firstly, what might cause a speaker to vary his/her use of smallwords, secondly, how might a learner actually acquire smallwords and, thirdly, what are the implications from what has emerged here for language education, with respect to both the way language is assessed and the way it is taught and learnt in schools.

Variation in smallword use

It is well documented, particularly in sociolinguistic studies, e.g. by Labov 1972, Cheshire 1998 and Milroy 1987, that language use varies according to both characteristics of the individual, such as gender, social class, age or 'personality', and characteristics of the situation, such as speaker relationships, degree of formality and the actual task being performed. Smallwords, although largely constrained by the signal(s) that they are used to send, are nonetheless likely to be subject to some variation due to these factors.

In the current study it has not been possible (owing to the relatively small size of the dataset and the nature of the tagging) to carry out an investigation into such variation in any breadth or depth. However, the fact that the data was tagged for gender and test task has made it possible to look for some cursory indications of the effect of these variables on smallword use.

Gender

Given that the quantities of smallwords produced by the three groups were not generally sufficient to allow more than tentative conclusions on the way smallwords were used to send signals, there could be no justification for further splitting this data to carry out an extensive analysis of these signals in relation to gender. However, there is a widely held view that females generally use more softeners and politeness forms than males (although for the

complexity of this issue see Holmes 1995, Coates 1986). It was therefore tempting to do a swift gender comparison of the smallwords explicitly associated with hedging in the dataset from the native speaker students. Here the male corpus consisted of 5,826 words, and the female of 11,803 words; in other words the male corpus was about half the size of the female.

Table 9.1 Comparison of use of some hedges and other 'politeness' terms by native-speaker boys and girls

	actual occurrences		**relative frequencies per 10,000 words**	
smallword	boys total: 5,826	girls total: 11,803	boys	girls
a bit	9	16	15.4	13.6
sort of	9	15	15.4	12.7
and things/everything/ stuff	7	14	12	11.8
please	11	21	18.9	17.8
thanks/thank you	12	21	20.6	17.8

As Table 9.1 shows, a comparison of the occurrences of *a bit, sort of* and *and things/everything/stuff* consistently showed the pattern that the boys produced approximately half the number of tokens that the girls produced, suggesting no relative difference. As a further test of politeness, the terms *please* and *thanks/thank you* (although not counted as smallwords) were also compared and, again, no relative gender difference was found. In fact, when relative frequencies per 10,000 words are considered, the boys are seen to very slightly outperform the girls in their use of all these terms. The initial indications are therefore that, at least in the case of smallwords associated with hedging, where a difference between the genders might have been expected, this did not occur. Clearly this could be explained by other sociological factors, given the small group of speakers, or by the unnaturalness of the test situation. This underlines the limitations of a small-scale study, where sociological representativeness of the subjects is not possible to engineer. As far as the Norwegian students' data was concerned, the tokens of hedging smallwords were too few for gender analysis. However, a look at the number of tokens of *please* and *thanks/thank you,* illustrated in Table 9.2 on page 226, suggests that, again, differences between the genders are slight, although here they are weighted in favour of girls. What is striking is the difference in quantities, particularly in the case of *please*, between the UK and Norwegian students, reflecting a cultural difference. In Norway it is not as common to use

a term equivalent to *please,* which probably accounts for the fact that, relatively, the Norwegian students produced less than a quarter of the *please*s that the NS students did. All in all, therefore, the tentative conclusion here seems to be that, while cultural background apparently has a significant effect on the use of markers of politeness, gender does not.

Table 9.2 Comparison of use of *please* and *thanks/thank you* by Norwegian boys and girls

	actual occurrences		**relative frequencies per 10,000 words**	
	boys total: 18,531	girls total: 16,373	boys	girls
please	7	7	3.8	4.9
thanks/thank you	22	21	11.9	12.8

Task

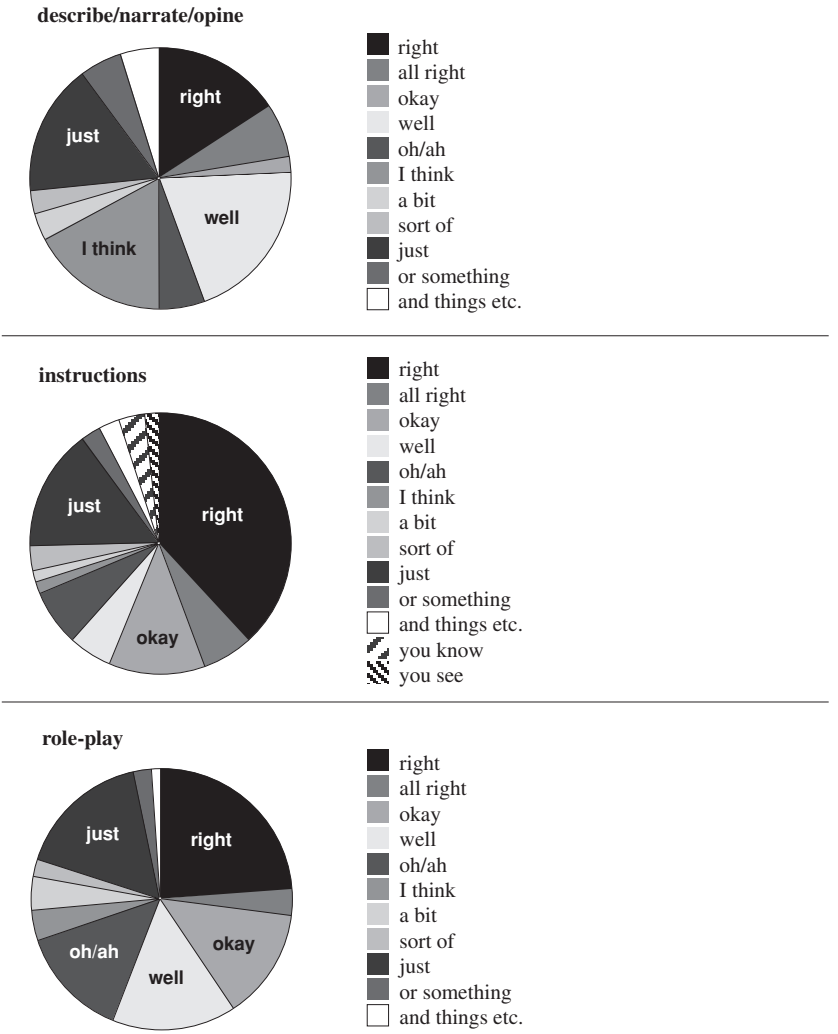
Clearly, the task performed by a speaker has a considerable effect on the language produced. This was taken into account in the design of the EVA test, and investigated in the content validation, as described in 'The content aspect of validity' pages 66 – 71. An analysis of the relative distribution of the smallwords used by the native speaker group across the three test tasks is illustrated in Figure 9.1. The most commonly used smallwords are labelled on the charts.

Task 1 involves describing a picture, narrating and discussing a topic associated with the picture sequence. Task 2 involves giving instructions, often in the face of difficult vocabulary, so that a considerable amount of checking, appealing and clarifying is involved. Task 3 is the role-play, and involves politeness to a greater extent than the other tasks.

As Figure 9.1 shows, task 1 particularly elicited *right, well, just* and *I think*, associated with structuring discourse and hedging opinions. Task 2 elicited many *right*s, as well as *okay*s and *just*s, a wide range of hedges, and a number of *all right*s, *you know*s and *you see*s, typically associated with the language of instructing and checking, appealing and clarifying. Task 3 elicited a relatively high number of *right*s, *okay*s, *just*s, *well*s, *oh*s and *ah*s, as well as a wide range of hedges, suggesting that role-play succeeds in heightening the sensitivity of speakers to the 'face' of the other.

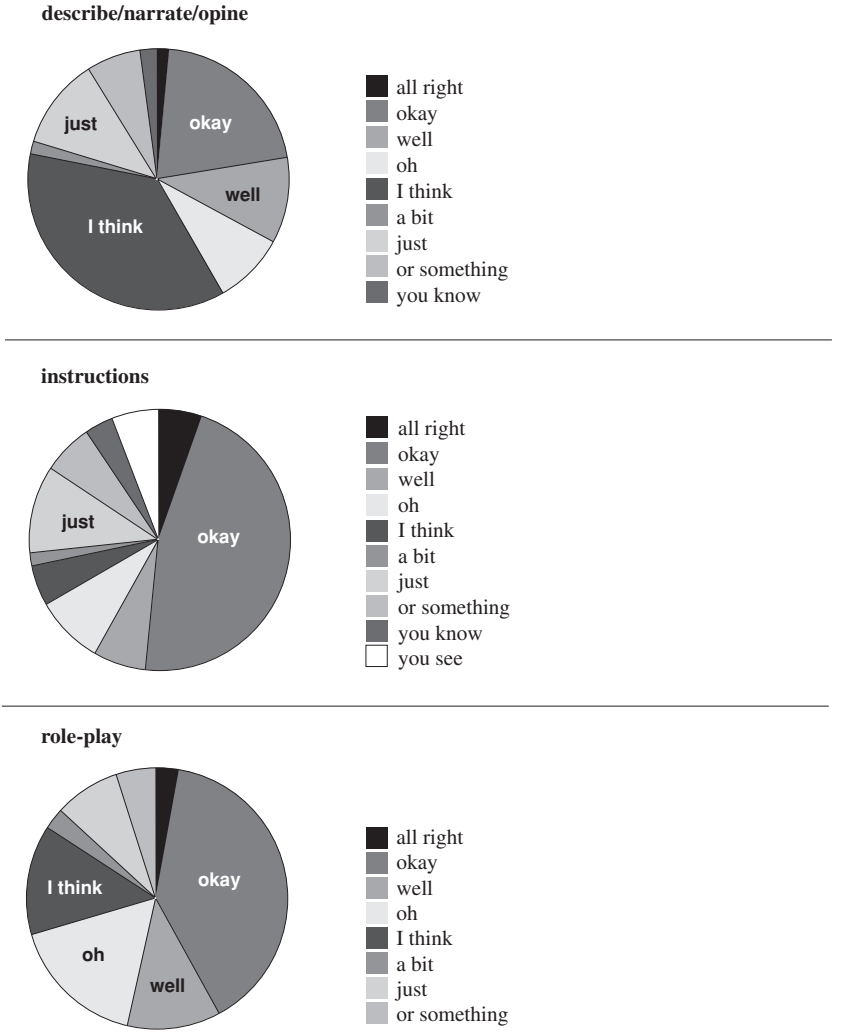
With the possible exception of *right*, in the case of instructions (where this word is used with both rising and falling intonation), it can be seen that no single smallword dominates within a task type and that there is a clear difference in the smallword choices across task types.

Figure 9.1 Distribution of smallwords across tasks: native-speaker group



In contrast, Figure 9.2 on page 228 shows the distribution of smallwords across tasks by the Norwegian group as a whole. Here, task 1 is marked by a dominance of *I think*s, mainly supported by *okay*s, *just*s and *well*s. Task 2 is heavily dominated by *okay,* as is task 3, which also elicited a certain number of *well*s and *oh*s, with hedges being mainly represented by *I think*.

Figure 9.2 Distribution of smallwords across tasks: Norwegian group



Here, the same tendency is reflected as in the case of smallwords generally, i.e. that the repertoire of smallwords used by Norwegians is narrower than that of native speakers, being confined to more 'central' terms, which tend to dominate within tasks. It is also noticeable that *okay* is given a role that reduces the variety across tasks, although even in the case of the Norwegians, the speakers do vary their smallword usage to some extent according to task.

This investigation, although being limited to what might be considered 'macrotasks' and not looking into the many 'microtasks' that are performed as speaking progresses, shows clearly that the native speakers and learners alike seem to feel the need to draw on different smallwords as the task type varies. This finding, coupled with the conclusions on learner usage above, highlights the value of varying tasks, in order to elicit and ultimately expand the range of smallwords at the speaker's disposal, however limited this may be.

The acquisition of smallwords

In Chapter 6, a case was made for a direct association between the use of 'chunks' in a foreign language and fluency in that language, citing the observations of Raupach (1984) and Towell *et al.* (1996). Learners were found to lack these chunks in their L2 prior to a stay abroad, but acquired them during that period, accompanied by a significant increase in measurable fluency. Moreover, it has been observed by Raupach that, in the case of chunks used as 'fillers' (which seem to correspond roughly to smallwords), the repertoire acquired by learners tends to be restricted and similar across learners (with shared L1s). Towell *et al.* offer an explanation as to why the use of chunks leads to greater fluency, arguing that chunks tend to be proceduralised language, used automatically and taking much of the formulation burden off the speaker in production. However, as yet, no explanation has been touched on of how learners acquire smallwords, and what might account for the selective way they appear to be acquired.

Although there is extensive literature on discourse markers and on second language acquisition, relatively little has been written about the combination of the two, although the upsurge in discourse research using learner corpora is beginning to remedy this, e.g. in Granger *et al.* (2002). In an attempt to shed some light on factors that may have been influential in the acquisition of smallwords in this study, this section will look at some findings that have been produced by other researchers on the acquisition of 'chunks' of language that seem broadly similar to smallwords, and at some SLA theory that relates to lexical items generally.

First, however, it is necessary to make clear just where smallwords seem to fit into the scheme of what are referred to as chunks, and whether they can justifiably be treated as lexical items for the purpose of the discussion. Nattinger and DeCarrico (1992) equate the expression 'lexical phrases' with 'chunks' of language, which they define as:

> multi-word lexical phenomena that exist somewhere between the traditional poles of lexicon and syntax, conventionalised form/function composites that occur more frequently and have more idiomatically determined meaning than language that is put together each time. (1992: 1)

The authors further analyse lexical phrases to include a category called 'polywords', defined as short phrases that function very much like individual lexical items, that can be both canonical (conforming to syntactic convention) or non-canonical, allow no variability (i.e. occur in one form only), and which are continuous (i.e. never being broken up by intervening words) (1992: 38). The authors add: 'Polywords are associated with a wide variety of functions, such as expressing speaker, qualification of the topic at hand, relating one topic to another, summarising, shifting topics and so on.'

The ensuing discussion clearly confirms that Nattinger and DeCarrico's polywords, being those lexical phrases situated at the most lexical end of the lexicon–syntax cline, provide a location for most smallwords as they are defined in this study. Those that fall outside the polyword location do so on the technicality that they are made of a single word, e.g. *right* or *well*. However these words share the characteristics of the polyword, insofar as they allow no variability. Whereas lexical phrases/polywords by definition exclude single words, this distinction has not been made when defining smallwords. This is partly because the boundaries here are fuzzy: a smallword can be made up of two words but written as one, e.g. *anyway*, or pronounced as one, e.g. '*kinda'* Besides, since polywords apparently function as single words, the distinction seems trivial in the context of this study. Following from Nattinger and DeCarrico's discussion, smallwords might thus be regarded as fixed, unvarying, continuous 'chunks' of language, consisting of one word or more, but functioning as single lexical items.

There is evidence from both L1 and L2 acquisition studies (e.g. Peters, 1983 and Wong-Fillmore 1976) that supports the idea that lexical acquisition is primary, insofar as learners first acquire unanalysed items of language as chunks, only later breaking these down into constituent syntactic or morphological parts. Whatever their fundamental bases, the most commonly cited current theories on both 'taught' and 'natural' language acquisition (e.g. Krashen 1982, Pienemann and Johnston 1987 and Towell and Hawkins 1996) generally share the basic tenet that in order to acquire a form it is necessary for the learner to be exposed to it in such a way that it can be understood and for him/her to be given the opportunity to use it in communication. Recently there has been a return to the conviction that, in some cases, attention to form is also necessary for complete acquisition in the taught context (e.g. Lightbown and Spada 1993).

Given that smallwords are chunks that are used very frequently one might thus expect them to be readily acquired by learners. The fact that learners have apparently been able to acquire a body of smallwords after a relatively short time in the target language country, yet had not acquired them during several years of language study, perhaps highlights a deficiency in traditional teaching. Smallwords are rarely drawn to learners' attention, often being cleansed from school texts, and learners are seldom actively

encouraged to use them, the implications of which will be returned to in the next section.

However, the facts that only a limited repertoire of smallwords appeared to be acquired by the learners in Raupach's study, and that these were generally the same across learners, do suggest that some smallwords are inherently more difficult to acquire than others, given a particular L1 background. What then can affect the ease of acquisition of a smallword?

Language transfer, or cross linguistic influence is held to be very pervasive in numerous studies, e.g. Odlin 1989. However, the view, represented by the early Contrastive Analysis Hypothesis (see Towell and Hawkins 1996), that items that are similar or literally translatable across languages will be easiest to acquire is now held to be oversimplistic. In Nikula's (1996) study of pragmatic force modifiers (to a large extent equivalent to smallwords here), she did find that learners of English tended to use modifiers that were directly translatable from their mother tongue (Finnish). (The same appears to be the case with the learners of French cited by Raupach). However, Nikula concludes that 'evidence of non-transfer was more apparent than that of transfer' (1996: 227). The 'behaviour' of smallword use was not transferred. While in their mother tongue the learners used a wide range of modifiers generally, including many implicit ones, they tended to stick to a narrower range of modifiers in the L2, and in particular to favour explicit ones. This led to an overreliance on explicit smallwords such as *I think*, and an avoidance of the more implicit, such as *you know*, or *well.*

Another factor that could affect learners' behaviour is what Kellerman (1983) terms the perceived translatability of idiomatic expressions into the L2, which he believes to be highly influenced by the transparency of the words in the expression. If these are used in their most core sense, the learners will more confidently transfer these expressions into the L2 (rightly or wrongly). This may account for certain smallwords, which might successfully have been transferred to the L2, in fact being lost.

The 'lexical gridding' of languages can also lead to limitations in a learner repertoire. Dagut (1977) studied the lexical errors of Hebrew-speaking learners of English, and concluded that learners tend to assume that an L2 item will occupy the same lexical space as whatever is perceived to be the corresponding L1 item. Smallwords tend to be used in very complex ways, and may have meaning on many planes simultaneously. The words *well* and *right* in their smallword sense intrude into spaces that their cognate equivalents in Norwegian, *vel* and *riktig* do not. This could potentially lead to a rejection or an unawareness of the versatility of these smallwords by Norwegian learners of English.

This phenomenon of what Dagut (1977) calls the 'incongruities of lexical gridding' may extend beyond semantic considerations: a smallword such as *sort of* or *a bit* may well be used (for reasons of face-saving or politeness) with

a pragmatic force in the L2, that it lacks in the L1. Moreover, socio-cultural factors themselves might need to be taken into account: what is regarded as face-threatening might differ between the L1 and L2 cultures (as has already been perceived in the case of the low use by Norwegians of *please,* in the previous section).

This discussion has attempted to highlight some of the factors that can influence the ease of acquisition of smallwords and to explain why these superficially simple words and phrases, so abundant in native-speaker language may be far from simple to acquire. Learners seem to pick them up abroad more easily than in school, suggesting that teachers must be more open to them. However, there seems no question that certain factors abound that make individual smallwords inherently less accessible to learners. De Cock *et al*. (1999), in their study of vagueness expressions (as fixed recurring phrases) in French learners of English, conclude that:

> the apparent inability of advanced EFL learners to master the use of vagueness expression has at least three possible causes: systematic differences in the way vagueness is expressed in their French mother tongue and in English; shortfalls in teaching (the use of vague language may be stigmatised); and finally, lack of contact with native speakers, a particular problem for EFL learners. (1999: 78)

Such comments, along with the findings of Nikula on pragmatic force modifiers and of Raupach on chunks, might be taken to apply equally to smallwords. Their acquisition is problematic; not only must opportunity be provided for learners to hear and produce them, but they also need to be given considerable attention, in keeping with their complex nature.

These conclusions are of considerable importance to this study. They help to explain and to put into context the findings here on the limited – and limiting – repertoire of smallwords in the mouths of even the more fluent Norwegian students. Moreover, they have clear implications for what goes on in language classrooms, where smallwords should be given attention in a way traditionally unthought of.

The implications of the findings for language education

This study of smallword use would be incomplete without a consideration of its educational implications, not only regarding language assessment, which was the starting point of this research, but also for teaching and learning in the language classroom.

Implications for assessment

The findings on smallword use have implications both for the design of tasks used in testing or informal assessment and for the instruments used in the actual assessment process.

The conclusions on smallword distribution across task type, cited in 'Variation in smallword use' pages 224 – 229, reinforced the contention that underlay much of the discussion in Part One, viz. that test tasks must allow candidates to demonstrate their ability across the full range of components of CLA. Failure on this count was shown, in 'A unified framework for validation' pages 28 – 31 and 'The validation process' pages 65 – 87, to affect content and substantive validity. In 'The content aspect of validity' pages 66 – 71, it was maintained that the design of the EVA test succeeded in taking account of this. Task 1 was intended particularly to elicit evidence of textual ability, through structuring discourse in longer turns, and through demanding a certain amount of pragmatic ability in 'softening' opinions. Task 2 was designed to elicit evidence of strategic ability, through the need to negotiate meaning as well as checking understanding and clarifying, and to some extent of pragmatic ability, e.g. through the need to preserve and maintain face when asking favours. Task 3 was largely designed to test pragmatic ability, i.e. whether the candidate had language conventions associated with certain situations, or was able to adopt an appropriate degree of politeness when addressing adult strangers.

The findings in 'Task' on page 226, on the variation of smallword use by the native-speaker students with task type, suggest that speakers use different smallwords in carrying out different tasks. Not only do these results indicate that the tasks appeared to elicit the type of language they intended, but they further show that different components of CLA seem to require different smallwords, or groups of these. This reinforces the need to vary tasks in test design to take account of CLA in full, since it seems that proficient speakers actually draw on distinct pools of language as they move from task to task.

The implications of the findings in Chapters 7 and 8, on the part played by smallwords in the language of groups spanning a range of fluency, are especially significant for the design of descriptors of ability, whether used in actual testing schemes or in more informal types of assessment, including self-assessment. The hypothesised stages of the acquisition of smallword signals, summed up in Table 8.9 on page 222, (along with the findings on smallword quantities and positions, as well as more mechanical features such as pausing and turn length), provide data-driven information of the kind Fulcher (1996) calls for as the basis of descriptors of ability.

These findings can be used to sum up, in terms of what have been identified here as markers of fluency, the speech of three groups of 14–15-year-olds

across a range of levels of fluency. The NoB level might be regarded as typifying a below average English speaker of this age in the Norwegian context, with the NoA level typifying an above average speaker. The NS level would not be attainable by most non-native speakers, although there is a steady growth in the number of bilingual students in Norwegian schools. It would certainly be possible to recognise levels in between each of these and realistic to add a lower level, where there is very little fluency. This could give the basis of a six-level scale.

Fulcher (1996) describes fluency at five bands of ability (summed up in 'Identifying elements of fluency' pages 126 – 132) largely in terms of the kind of pausing (for content planning or language searching), and how this affects speaking. He considers the message itself – how expansively and confidently it is delivered – and the use of vagueness markers and repair and clarification. The information emerging from this study could be combined with Fulcher's descriptors to put together assessment instruments to suit different purposes.

To illustrate this a scale is drawn up below, designed primarily for rating purposes, which combines the findings from this study with some of Fulcher's findings, assuming that the extreme points of his five-band scale correspond to NoB and NS levels cited here. The result is three levels on a scale, which one can assume are preceded by a 'pre-fluency' level 1, and intermediary levels between those shown here, where level 2 corresponds to NoB, level 4 to NoA and level 6 to NS. Although smallwords are mentioned as examples of the kind of signal used, the descriptors are primarily couched in terms of the aspects of the communication that are affected by the fluency level of the speaker:

- length of utterance, pausing and speed
- textual links within turns
- interactive links
- repair, clarification, etc.
- marking vagueness, uncertainty
- pragmatic signalling – face saving, politeness.

(It is worth noting how closely these aspects correspond with what were listed in 'Speaking' pages 43 – 46, as the skills specific to speaking, largely adapted from Bygate 1987.)

Level 2 (NoB)

Speech is broken up into noticeably shorter stretches than native speakers use, with many pauses due to language difficulties rather than planning content. Speakers have a very limited capacity for explicitly signalling how what they say relates to what has been said before, either across/between turns or within their own speech; only a few simple, straightforward linking devices tend to

be used, such as *okay* and *then*. They are not adept at getting clarification or at signalling self repair. They have a very small repertoire of signals of vagueness and uncertainty, e.g. *I think*, which tend to be used in a literal, propositional way rather than for pragmatic effect.

Level 4 (NoA)

Speakers are able to produce stretches of speech that are not noticeably curtailed by language difficulties. A number of pauses when searching for words slow down the speech to some extent. Although they do this less frequently and with a smaller repertoire than native speakers, they explicitly relate what they say to what has just been said, even at times when this relationship is less than straightforward, e.g. when there is unexpectedness or discontinuity, often signalled by *well*. They are as likely as native speakers, (although lacking their range) to give backchannels, such as *okay* to other speakers and to acknowledge what is said; however, they use appealers, like *you know* rather less. They signal self-repair. They use a limited range of vagueness markers, such as *just,* with a rather neutral pragmatic effect.

Level 6 (NS)

Speakers at this level produce utterances that are rarely broken up or slowed down by pausing other than for planning purposes. They have a wide range of devices for linking speech across and within turns, and complex relationships, such as returning from digressions, e.g. *well anyway*; these are selected for pragmatic as well as cohesive effect. They maintain the flow through backchannels and acknowledgement of what the other speaker says, and counteract possible breakdowns through checks and clarifications. They have a wide range of vagueness markers, such as *a bit* and *and things like that*, which are generally used for pragmatic effect rather than for propositional uncertainty. They are able to exploit the pragmatic force of a wide range of markers that allow them to signal empathy, e.g. through *you know,* and pleasure, e.g. through *ah*.

Another way of using this kind of detailed information on fluency is to emulate the recent work carried out, e.g. by the Council of Europe, whereby levels are broken down into a number of 'can-do' statements, which can be used in self-assessment, or for teacher observation (see Council of Europe 2001). The advantages of using such statements are considerable. The individual learner is rarely exactly 'at' a particular level, normally having strengths and weaknesses. By considering a range of statements, roughly corresponding to the aspects listed above, s/he will be able to assess where these strengths and weaknesses lie, and will be able to set concrete objectives,

and to track progress over a shorter period of time than it would take to move from level to level. Moreover, the statements are couched in entirely positive terms, only stating what the learner can do.

An example of how such statements could be made for self-assessment, associated with the three levels shown above, might be as follows:

Level 2

I can keep speaking for a short time and link my ideas together in a straightforward way, e.g. with *and, but* or *then*.

I can show that I have understood what is said, and that I am ready to take the turn, e.g. using *okay*.

I am able to show that I am not sure of what I am saying, e.g. using *I think* or *or something …*

I am able to show that I am surprised by what the other speaker said, e.g. using *oh.*

Level 4

I am normally able to keep going long enough to say what I want to, linking the parts of what I say, even if at times there is a break in the logical flow, e.g. using *well*.

I am able to encourage the other speaker to keep going, e.g. using *okay* or *I know*.

When I take the turn I am able to signal that what I am saying is not quite what the other person expected or hoped, e.g. using *well.*

I am able to signal that I am going to correct what I have just said, e.g. using *I mean* or *well.*

I am able to check that the other speaker is following what I am saying, e.g. using *you know*.

Level 6

I normally only pause to collect my thoughts and to plan what I want to say, rather than because I am searching for the right way to say it.

I am able to keep going even if the logical flow is interrupted in different ways, such as by changing the subject, digressing, going back to earlier topics, e.g. using *by the way* or *anyway.*

I am able to use a variety of ways of showing that I wish to take, hold or yield the turn, and I can combine these, e.g. using *right, okay, well …*

I am able to use a variety of ways of checking that the other speakers are following what I am saying, e.g. using *you see, you know* or *right?*

I am able to use a variety of ways of clarifying what I mean, e.g. *I mean, what I mean is* or *in other words…*

I am able to use a variety of ways of showing surprise, so that I can also show how I react to what is said, e.g. *oh I see, ah* or *oh well.*

I am able to use a variety of ways of 'softening' the effect of what I say, so that I don't sound too self-assured, e.g. *sort of* or *a bit.*

I am able to show that what I say covers a lot of things, e.g. using *and things like that* or *and everything.*

Clearly there are many ways of using statements such as these. For instance, the partial achievement of those associated with level four might indicate a transitionary level three. Or they need not be directly associated with levels, but rather be regarded as reflecting a gradual progression in the personal acquisition of fluency. They may be added to or further broken down, e.g. with separate statements for digression, interruption and so on. However they are used, they will serve as an awareness-raiser and a progress-tracker of the kind of things a learner needs to be able to do in order to become more fluent, and the sort of language he /she needs to acquire in the process.

Implications for teaching and learning

The implications of the findings here for teaching and learning stem from three principal conclusions: firstly that there is a need to vary spoken tasks; secondly that smallwords are necessary for fluency, and thirdly that these may be difficult to acquire and need to be worked on consciously.

The need for tasks to fully activate CLA is, if anything, even more important in the learning situation than in the test situation – what actually goes on in the language class can be argued to have greater long term consequences for learning than what goes on in a test. While most courses cover a range of topics, and many consciously present learners with the language necessary for a range of language functions, the actual tasks that learners routinely perform do not always give them the opportunity to exercise the full range of the ability they are supposed to be developing. Typically, the oral part of a lesson might consist of asking and answering questions with a classmate, or giving factual information. The focus is frequently on using the 'piece of language' that is being taught, thus primarily building up microlinguistic ability.

It is therefore important that other components of CLA should not be neglected in classroom interaction. For textual ability to be built up, the

learner should get the opportunity to speak in longer stretches as well as in shorter-turn exchanges. They should also get the opportunity to practise discourse with particular structures, e.g. when using the telephone. Developing pragmatic ability requires the opportunity to adapt speech to a range of conditions, as well as adopting the conventional ways of performing a range of functions. Strategic ability grows with learning how to cope with problematic communication, e.g. where vocabulary is unlikely to be known by all parties.

Those involved in classroom activity – coursebook writers, syllabus planners and teachers, and to some extent, learners themselves – should be aware of the importance of ensuring that the tasks carried out cover these needs. Learners should be systematically put into a range of situations that activate and develop skills fully. Role-play and drama are particularly valuable for widening the scope of what is talked about, what is 'done' with the language and the conditions under which it is spoken. This can be achieved to some extent by a simple manipulation of a classroom exercise, whereby students cease to be 'themselves in the classroom' but act as someone else in another place; even giving students English names for the English lessons has been found to have some effect in liberating them from the immediate situation. Table 4.1 on page 62, designed for evaluating test tasks can be used equally well as a framework for judging the interaction in the language classroom itself.

The implications for teaching of the findings specific to smallwords are possibly more complex, since these are a matter of influencing not simply the substance of teaching, but also the attitudes behind it. As was noted in the previous section, smallwords have traditionally been neglected in the language classroom, partly through genuine difficulties involved in teaching them, but largely through their low status. This status is probably something to do with the fact that they are not generally found in the written language, and the written word has traditionally enjoyed a higher academic status than the spoken, being the basis of literature, study, examinations, etc., besides forming the source of most texts presented to learners. Moreover, where smallwords are referred to in the literature on discourse, it has been normal until fairly recently for them to be simply grouped under the general term 'fillers', together with *erm* (e.g. Brown and Yule 1983b: 17). There is still a tendency for non-native teachers to regard the use of these as a weakness, and not something they would encourage in their students. And dialogues in coursebooks still tend to be cleansed of many of the very words and phrases that characterise living dialogue. Only when presented with evidence that it is the native speakers who really do use these, with learners lagging well behind, are teachers normally convinced that this is a body of language their students need to learn.

The traditional dearth of smallword teaching may also be due to the fact that they present certain difficulties as teaching matter. One problem is the fact

that they have been difficult to come across in the kind of texts normally used in the teaching context. Another is that, as we have seen in 'The acquisition of smallwords' pages 229 – 232, their multifunctionality makes them unlikely to have absolute equivalents in an L1. A third is that they are difficult to pin down regarding a rule or explanation of how they are used. However, none of these problems is, I believe, insuperable today. The availability of language corpora, the widespread use of tapes/CDs with scripts to accompany coursebooks, the ease of video recording, and sound-bites on the Internet all facilitate the study of authentic spoken language in a way undreamt of a decade or two ago. Nor should the absence of explicit 'rules' about how smallwords work continue to pose an obstacle. Language study has become less prescriptive, and more descriptive and, since the advent of Schiffrin's (1987) classical study of discourse markers, there has been a surge of interest in the description of the role played by this body of language in spoken interaction, e.g. in Stenström (1994).

One of the great advantages of smallwords as subjects for learning is their ubiquity. While texts have often had to be artificially constructed to provide examples of vocabulary or grammatical structures, the same cannot usually be said of smallwords, which are everywhere (although not randomly). Students can be given awareness-raising activities, using video and task sheets; they can also study tapescripts or downloads from corpora, being asked to notice which smallwords are used when, and why, or to guess which smallword has occurred (gap-filling), or predict what may follow one. Nolasco and Arthur (1987) suggest many such types of activity, and Dörnyei and Thurrell (1992) provide teachers with techniques for helping learners to acquire the smallwords and other words and phrases that bring interaction to life, in a way that few coursebooks normally attempt. Thus there seems to be reason for optimism that the teaching of the spoken language is poised to benefit from both the ability and the will to put smallwords squarely on the agenda.

Summary

This chapter has shifted the focus from smallwords themselves onto more human considerations of the person who is – or should be – using them. Three main issues have been addressed: factors that cause smallword use to vary, the acquisition of smallwords, and the implications of the findings from this study of smallwords for educational practice.

A thorough study of variation in smallword use can only be carried out by sociolinguistic research, using appropriate datasets. Within the limits set by the present dataset however, it has been possible to look for some indications that gender or task might affect smallword use. No clear signs were found that gender influenced the smallword use of the students in this study (although

there was some suggestion of cultural influence). There is clearly a need for more extensive study of the influence of gender and other learner characteristics on smallword use. There was, however, significant evidence that task did have an effect, in the case of both native speakers and learners.

In the absence of a significant body of research into this immediate area, the section on the acquisition of smallwords drew heavily on research into vocabulary acquisition, as well as on studies in the role of language 'chunks' in the onset of learner fluency. It was concluded that smallwords, because of their complexity and multifunctionality, tend to belong inherently to those vocabulary items that cause problems to learners. However, their commonness should compensate for this, but only if learners are allowed to be actively exposed to them and made aware of them, which in turn puts the onus on the teacher/material writer to accept and acknowledge their significance.

The implications for teaching, learning and assessment were found to be many. The findings on smallword use throughout this study have built a strong case for a new focus on this body of language in both the assessment and the acquisition of fluency. It has moreover been demonstrated that only a varied diet of tasks/functions enables a learner to use smallwords widely and discover the need to extend his/her repertoire. Examples have been given of how the findings of the study might be utilised in writing descriptors of language ability, with both reporting and pedagogical value. And practical ideas have been offered for breaking the mould by placing smallwords firmly on the agenda in the language classroom, where they have hitherto been undervalued or ignored.

Conclusion

10 Conclusion

This book has given an account of a complex piece of research, which was complex for several reasons. Firstly, it had a double focus – it was as much about learner language, specifically smallwords and their role in fluency, as it was about test validation. Secondly, frameworks had to be devised for the analyses, which drew on very diverse theoretical fields, bringing in Messick's (1996) theory of unified validity and Sperber and Wilson's (1995) relevance theory, as well as discourse analysis and second-language acquisition theory. Thirdly, the empirical research drew on a wide range of data types, with ratings, biodata, diverse judgements and self-assessments on one hand and learner and native-speaker corpora on the other. Finally, it was carried out by myself wearing a number of hats. I was heavily involved in the designing and rating as well as being the validator for the particular test in question, and was motivated equally as tester and teacher to carry out the research for its more generalisable findings.

It is not the intention in this conclusion to reiterate all that has come out of the complexity of questions and answers that comprise the study. However, certain pivotal research questions were laid down in 'Research questions' pages 3 – 4, and it is worth recapping these and briefly summing up the findings that emerged, as the study attempted to answer them.

The research questions

Seven empirical questions were identified on pages 3 – 4, as follows:

- 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 corpus evidence of non-linguistic, temporal features which supports the score-based grouping of students into more and less fluent speakers?
- 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?

In order to address these questions, a number of theoretical questions had to be considered, which involved scouring the literature rather than analysing data. These might be summed up as follows:

- How might we systematically test the validity of a test of spoken interaction?
- What exactly do we mean by communicative language ability (CLA), in the context in question?
- What do we mean by fluency, and is there any suggestion in the literature that smallwords and fluency go hand in hand?
- How might we systematically analyse smallword use?
- What do we know about the acquisition of smallwords?

The findings

This section will consider the research questions in turn – first the theoretical, then the empirical. A brief account will be presented of the principal findings relating to each question.

Theoretical findings

How might we systematically test the validity of a test of spoken interaction?

Working out a systematic and comprehensive way of validating a test such as the EVA speaking test was the subject of Chapter 2. Indeed the findings of that chapter largely relate to any language test, although the final framework is couched in terms that particularly apply to the test in hand. Although the literature on language-test validation was found to provide a detailed and comprehensive picture of test validity, there was inconsistency among writers regarding both which types of validity to include and how widely to define these various types. The most systematic account was found in Messick (e.g. 1996) where validity is seen as a unified concept but regarded as having six aspects, which may overlap but which between them cover the whole 'thing'. This, and the fact that Messick's account of validity has been granted considerably currency in the language-testing community worldwide, lay behind the decision here to build a six-part framework for validation on Messick's account.

As a preliminary step, nine types of validity were considered, which seemed to broadly cover validity as it is presented in more recent literature

(e.g. Alderson *et al.* 1995), and, for each of these, the major potential threats to validity in a test of the type studied here were listed (see 'Threats to validity summarised' pages 27 – 28). Each of these sources of invalidity, of which there were over 30 altogether, was then assigned to the aspect(s) of validity in the six-part framework which it was judged to threaten directly. Thus a system was created whereby individual, very concrete sources of invalidity could be investigated in turn, and their potential effect on the test's validity could be localised to one or more of the following aspect of validity, following Messick (1996):

- CONTENT
- SUBSTANTIVE
- STRUCTURAL
- GENERALISABILITY
- EXTERNAL
- CONSEQUENTIAL.

For definitions of these aspects, as well as an overview of how individual sources of invalidity were identified as threats to the particular aspects, see 'Six central aspects of validity' pages 29 – 30.

What exactly do we mean by communicative language ability (CLA), in the context in question?

The discussion of what goes into CLA, covered in Chapter 3, is traced back to Hymes (1972), who proposed the then-revolutionary idea that being able to communicate in a foreign language depended not only on a knowledge of the basic structures of a language but also on knowing what it is 'normal' and appropriate to use at any time, as well as being able to put this knowledge into practice in actual performance under real-world conditions.

Building on Hymes, a number of models for CLA have since evolved, generally with between three and five basic component labels, but maintaining a fairly high degree of consensus regarding what is essentially involved in CLA. On the basis of this, a four-component model of CLA was proposed here. The components, which are defined in 'A suitable model of CLA' pages 39 – 42, are:

- MICROLINGUISTIC ABILITY
- TEXTUAL ABILITY
- PRAGMATIC ABILITY
- STRATEGIC ABILITY.

In order to operationalise, or state in concrete terms, what we expect this ability to require of our students when taking part in spoken interaction, it was necessary to consider the particular demands imposed by spoken interaction itself, as well as the various conditions under which speaking would be likely to take place. The former largely involved a consideration of Bygate's (1987)

skills associated with speaking, while the latter took into account Weir's (1993) four variables: purpose, interlocutor, setting and channel, as well as topics, functions and level of ability. Clearly the school curriculum played an important role in defining the scope of these.

The resulting model of CLA lists in some detail what students would be expected to be able to do, if they had acquired optimal CLA in all its aspects. The model, which can be seen in 'Summary' pages 55 – 57, has the advantage that it helps students and teachers to see fully what can be expected of learners, and where weaknesses exist, it casts light on how these affect performance. Like the validation model, it is comprehensive yet not made of watertight parts; there is interdependence and overlap in both models.

What do we mean by fluency, and is there any suggestion in the literature that smallwords and fluency go hand in hand?

The question of what 'fluency' means is a puzzler, as has been proved by researchers (e.g. Esser 1996) who have taken on the arduous task of pinning it down. Everyone 'knows what it is' yet few feel competent to define it and there is only partial consensus among those who do. However there was found to be sufficient agreement on what is at the core of fluency to arrive at a working definition, in Chapter 6, as:

> the ability to contribute to what a listener, proficient in the language, would normally perceive as coherent speech, which can be understood without undue strain, and is carried out at a comfortable pace, not being disjointed, or disrupted by excessive hesitation.

Literature was consulted in the attempt to link smallwords and fluency, both as a secondary source of empirical research and to establish theoretical grounds. Traditionally, empirical studies of fluency (in common with test descriptors) tend to have focused on what are here termed *temporal markers* of fluency, such as:

- increased overall rate of speech
- increased mean length of unbroken run of speech
- decreased frequency of disruptive unfilled or non-verbal filled pauses.

However, a major aim in this study was to identify markers of fluency that are helpful to the learner, who needs to know if there is any particular language that will build up his/her fluency. Studies that actually analyse learner language at varying fluency levels, such as Towell *et al.* (1996), were found to suggest that the following linguistic features are associated with fluency:

- increased nativelike use of formulaic expressions generally
- increased nativelike use of smallwords.

The theoretical link between fluency and smallwords is the subject of the second part of Chapter 6. The literature on smallwords (under other names,

e.g. in Schiffrin 1987) suggests explicitly that the functions they perform facilitate what is described as fluent speech in the working definition used here. Furthermore, this definition is highly compatible with the relevance theory of communication, which is all about bringing about coherence and comprehension with a minimum of strain (and words). This, according to Sperber and Wilson (1995), is dependent on a hearer's being able to draw inferences, which in turn depends on certain parameters being set. It is argued here that smallwords send the cues, or signals, necessary to set these parameters and thus to bring about fluency.

How might we systematically analyse smallword use?

The relevance theory explanation of the role of smallwords is the basis for the much needed framework for analysing smallwords in use. Five major parameters need to be set for communication to run smoothly, according to the interpretation of relevance theory in this study. These parameters, which are fully defined in 'The work of smallwords in optimalising fluency' pages 142 – 148, are:

- the communicative intention of the speaker
- the context for interpretation of the utterance
- the cognitive effect of the previous utterance
- the degree of vagueness or commitment in an utterance
- the state of success of the communication.

For each of these parameters several settings can be made, and it is claimed here that it is the role of smallwords to point the hearer towards these individual settings. In the framework for analysing smallword signals, the five parameters are classed as macrosignals, while the individual signals sent by smallwords to set these, are termed (micro)signals. The framework is shown fully in Figure 6.2 on page 147. Signals can be assigned to smallwords on the basis of pre-defined contextual slots or, in a minority of cases, on the basis of inherent meaning ('Defining and analysing evidence that smallwords are used to send signals' pages 188 – 218).

The need for such a framework is particularly acute because studies of smallwords either tend to assume native-speaker competence (and thus have less need for contrasting the different ways smallwords are used) or they focus on individual smallwords or groups of these at a time. The framework worked out here provides a unified and comprehensive way of analysing any smallword in context. It is based on the assumption that smallwords are used as a system by the speaker, who will draw on his/her own stock, however limited, and use these to send signals as well as s/he is able (thus ensuring that certain 'safe' smallwords may be given extended signalling powers at the expense of other, less familiar ones).

What do we know about the acquisition of smallwords?
The literature that explicitly addresses the acquisition of smallwords seems to consist of relatively few empirical studies on a limited number of smallwords, such as that of Nikula (1996). However, there is a considerable body of literature on the acquisition of vocabulary and on lexical 'chunks', of which smallwords seem to be a group, and in Chapter 9, this literature is turned to.

There is a consensus that frequency of encounter is a factor in facilitating acquisition of an item (e.g. Krashen 1982). Moreover, there is evidence in both L1 and L2 research that 'chunks' lend themselves well to being acquired as such, and used automatically (see Nattinger and DeCarrico 1992 and Towell *et al.* 1996). Why, then, is there evidence, both in this study and in others (e.g. Raupach 1984), that learners fail to acquire some of the most common smallwords, at least until after a period of residence in the target-language country?

Part of the explanation for this is derived here from the literature on vocabulary acquisition, where a variety of reasons are offered as to why some items are more difficult to acquire than others. It seems, for example, that 'opaque' idiomatic chunks – such as smallwords tend to be – are often regarded as untranslatable (see Kellerman 1983, backed up by Nikula's (1996) conclusions on implicit pragmatic markers). Moreover, the different ways in which we divide our lexical spaces up can be an obstacle in moving from L1 to L2, according to Dagut (1977), and this could be significant in the case of smallwords that can operate on several planes simultaneously; it is unlikely that smallwords in one language would occupy the identical spaces to a corresponding item in another. Thus it seems that the nature of smallwords may make them difficult to acquire despite their 'smallness' and frequency.

However, it seems that there is more to it than that. The fact that learners appear to acquire these readily once they are abroad suggests that the environment in the foreign-language classroom back home is somehow not conductive to smallword acquisition. This feeling is backed up by De Cock *et al.* (1999) in their findings on 'vague' lexical phrases, which they believe are neglected and even stigmatised in teaching. Traditionally this body of language has not been given attention in a world where the written text has dominated and where even so-called spoken texts tend to be 'purified'. Clearly there needs to be a change in attitude to smallwords, with a recognition of their significance, which studies of this kind may go some way towards fostering.

Empirical findings

Which aspects of validity appear to be at risk in the test 'as it stands', and what are the likely causes of invalidity?
In Chapter 4, the test was scrutinised for each aspect of invalidity, checking

each of the potential sources of invalidity in the framework. At this stage, only the test 'as it stood' – i.e. the material used for the test itself, including the rating instruments and all instructions – was investigated. This meant that no judgement could be made on the way the test was actually functioning; it was only possible to look for flaws in the design which could actually prevent it from working properly.

The overall conclusion was that the material used in the actual test (e.g. the tasks, the format and the instructions) was designed well and did not in itself pose a threat to validity. The rating instruments, however, were less satisfactory. This result was unsurprising; a great deal of work had gone into the design of the tests, of which there were three versions. They involved a variety of tasks – as authentic as possible – based largely on entertaining pictures and dealing with relevant situations, which were designed to cover CLA as it was operationalised here. The students carried out the tasks in pairs, with clearly defined roles, maintained through the use of a script for the tester. Experts were involved throughout the design process and trialling was carried out in several rounds, with teachers involved at every stage, providing judgements and feedback on the test material. Although the rating instruments had been carefully designed, e.g. with reference made to other rating systems, teachers consulted for estimates of students' ability and video guidelines issued for raters, the emphasis had been on the format of the rating and reaching agreement on overall levels, rather than on achieving some kind of 'match' between items on the band scales and profiles and the components of CLA in the operationalised model.

In fact, a major weakness was found to lie in a mismatch between the instruments and the model of CLA, with strategic and textual components being poorly represented, reflected in gaps in the band scale associated with fluency. Moreover, there was perceived to be some vagueness in the language in the instruments, and the lack of student self-assessment was felt to be a shortcoming in the testing process. These flaws led to an undermining of the test's validity in its STRUCTURAL, GENERALISABLE and CONSEQUENTIAL aspects. Table 4.5 on page 94, a profile of validity, shows the findings from this stage in the validation process, indicating the actual weaknesses found as well as other areas where empirical evidence is needed in order to make judgements.

How far do raters' scores provide actual evidence of this suspected invalidity, and shed further light on its causes?

The aspects where raters' scores were able to add to the findings on validity were identified in Chapter 5, as:

- the EXTERNAL aspect (investigating the degree of correlation between scores and other measures of ability)

- the STRUCTURAL aspect (using factor analysis to find support for the way sub-skills were grouped on the rating instruments
- the GENERALISABLE aspect (looking for evidence of gender bias, rating inconsistencies and judgements of vagueness in the wording of instruments).

The data from teachers' estimates of ability and from students' self-assessment (carried out independently of the testing) correlated sufficiently well with the overall test grades, when compared with other similar studies, to permit the conclusion that the EXTERNAL aspect of validity did not appear to be flawed. Evidence of divergence with tests of other skills further supported this.

Factor analysis of sub-skill scores on the performance profiles gave support to the way language and fluency were separately rated on the band scales, with pronunciation and intonation judged independently. This was a positive result for the STRUCTURAL aspect of the test's validity.

Although gender bias was not detected in the test scores, when these were compared with teacher estimates, inter-rater reliability was found to be poor, particularly in the rating on certain sub-skills – notably those associated with fluency. This was felt to be bound up with a fairly low judgement of clarity in the wording of the instruments. Thus the test was found to be in need of some improvement in rating instruments and processes in order that the GENERALISABILITY of its scores could be considered valid.

The results of the investigation using scoring data are shown in the reappraised profile of validity, Table 5.11 page 116. Here a number of issues were outlined for future research – these concerned virtually all aspects of validity, and were found to be dependent on data that could only be collected consciously during or after testing, from students and teachers/testers or from others affected by the use of the test.

However, one area was earmarked for further investigation within this study, using the language corpora compiled during the trialling process. This concerned the STRUCTURAL aspect of validity, where clear weaknesses were detected. There was found to be too little reference to concrete evidence of strategic and textual components of CLA as these have been operationalised here. Furthermore, and hand in hand with this, the band scale associated with fluency was found to lack reference to actual linguistic performance.

It was deemed that any items being introduced in order to compensate for these deficiencies in the rating instruments should be concrete and easily recognisable (in the interest of the GENERALISABILITY and CONSEQUENTIAL aspects), and should be linguistic, rather than temporal (in the interest of the CONSEQUENTIAL aspect). Smallwords were proposed, as a body, as the candidates with greatest potential on all these counts.

Is there corpus evidence of non-linguistic, temporal features which supports the score-based grouping of students into more and less fluent speakers?

The study turned next to looking at the language of students at different levels of fluency, ultimately to compare smallword use. This involved the use, in Chapter 7, of electronic corpora, based on the transcripts of Norwegian students (tagged for overall test grade) as well as those of native-speaker students. Two groups of Norwegian students were defined as more (NoA) and less (NoB) fluent, sorted by overall grades (worked out from levels on both the language and fluency scales). A third, control, group of native speakers (NS) was also defined. The three groups were designed to be roughly similar in size and to be balanced according to gender. Before proceeding with a quantitative investigation of the smallword use of the groups, it was necessary to establish that they were indeed differentiated by fluency, measured independently.

Two measures of fluency were used: filled pausing and mean length of turn, with reference to research on fluency reported in the literature, e.g. Freed (1995) and Lennon (1990). The results of chi-square tests were interesting, but reassuring. In all cases, student groups defined as less fluent had shorter turns and relatively more mid-utterance filled pauses than those defined as more fluent. Initial pausing – presumably necessary to collect thoughts – did not differentiate the groups. What can be concluded is that, judged in terms of what Fulcher (1996) regards as 'disruptive pausing' and hence length of unbroken speech (and hence speed, following Towell *et al.* (1996)), as well as in mean turn lengths, there were unambiguous indications that the NoB students were less fluent than the NoA students, who, in turn, were (not surprisingly) less fluent than the NS students.

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?

Smallwords were counted and contrasted between the three student groups (NoB, NoA, NS) in a number of ways as Chapter 7 proceeded. Because smallwords play such a significant role in interaction, and because part of textual competence as it is defined here implies the ability to create cohesion both within and across turns, it was decided to compare smallword use in different turn positions: turn-initial, turn-internal, turn-final and 'loner' (i.e. without actually taking the turn, e.g. as a backchannel). Moreover, as it was important to study how the groups varied in their smallword selections, it was decided to investigate the range of different smallwords used by the groups. As in the case of pauses, relative quantities (to total words or turns as appropriate) were studied, due to some discrepancy in the amounts of speech uttered by the three groups.

The results were more or less as expected. Apart from turn-final

smallwords, which were too few to study, and loners, where the NoA group equalled the NS group, the tendency was clearly NoB<NoA<NS when it came to the numbers (overall and in different turn positions) and ranges of smallwords used. Moreover, it was found that smallwords used regularly by any group were always used by the more fluent group(s), but the reverse was not true. These results were encouraging for the study, but still begged the question of how the smallwords were being used.

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?

The question of how smallwords were used was addressed in Chapter 8. Here, the framework worked out in Chapter 6, and outlined in 'Theoretical findings' pages 244 – 248, was implemented. Every smallword in the dataset (consisting of a printout of all smallwords in context) was manually assigned, by agreement between two native speakers, to one or more signals, using contextual slot definitions in most cases. The numbers of smallwords used to send each signal and the actual smallwords chosen to send them were contrasted across the groups. The native-speaker students' usage was regarded as the yardstick, but since numbers were low in the case of some signals, and could be idiosyncratic, the literature on smallword use by native speakers was consulted as a back-up; a high level of concurrence was found in all cases.

The result, summarised in Table 8.8 on page 219, showed a clear patterning: the more fluent Norwegian students more closely resembled native speakers in the way they used smallwords to send signals than did the less fluent; this applied to the range of signals sent as well as to the choice of smallwords for each signal. However there were noticeable 'gaps' even in the speech of NoA students, who were very inclined to stick to a handful of seemingly more versatile smallwords; this was at the expense not only of the basic 'fluency' signals sent by some very common smallwords, such as *right*, but also of many of the pragmatic functions smallwords perform. This analysis, along with that in Chapter 7, corroborated the contention here that smallwords and fluency do seem to go hand in hand, and that even the stronger students noticeably lack finesse in using these. Table 8.9 on page 222, shows hypothesised stages at which smallwords appear to emerge, and provides a basis for further investigation into the acquisition of smallwords by learners.

How might these findings be applied to the assessment of fluency?

The application of the findings from this study to assessment is addressed in Chapter 9. The study has clear implications for the way fluency is described, and hence judged, in testing, through band scales or other rating instruments. However it also has implications for the tasks used in the assessment process.

The tasks in the current test were established, in Chapter 4, as having been

designed to cover the full range of components of CLA operationalised here, as well as the macrofunctions identified. This was largely corroborated by teacher judgements, and further reinforced by the distribution of smallwords across tasks by the three student groups, reported in 'Task' on page 226. The way the native speaker students clustered smallwords according to task was revealing, as it reflected the fact that different tasks demand different basic signals, and that these can be tempered by pragmatic considerations. The Norwegian students also varied their smallwords across tasks, showing that they too were yielding to the basic task demands; however the tendency (especially among the least fluent students) to stick to favourites (with peaks rather than clusters) highlighted their lack of ability to cope with the nuances imposed within task types. In fact, smallwords might be uniquely regarded as a barometer for what is actually demanded by any task. The range of smallwords used by students in our tests can be interpreted as symptomatic either of their language ability, or of the tests themselves, or of both.

In 'Implications for assessment' pages 233 – 237, the findings on the order in which smallword signalling appears to be acquired was drawn on and combined with the quantitative findings from Chapter 7, on both temporal markers and smallwords. This in turn was linked to Fulcher's (1996) findings on the language of learners across a range of levels, giving rise to a set of three descriptors, at increasing stages of fluency, primarily intended for use by raters. While these are exemplified by specific reference to smallwords, their focus is primarily on the communication itself, and how it is affected by the level of fluency, in terms of:

- length of utterance, pausing and speed
- textual links within turns
- interactive links
- repair, clarification, etc.
- marking vagueness, uncertainty
- pragmatic signalling – face saving, politeness.

As self-assessment has been judged here to be a crucial element in any (at least informal) assessment process, another set of descriptors – of the *I can do* type – was drafted, drawing on the same elements; this can also be seen on pages 233 – 237.

Can these findings ultimately be applied to raise the level of fluency in learners?

Ultimately, this study has successful learning at heart. This is reflected in the kind of assessment processes that are the subject of much of this study. However the findings also have direct implications for the actual learning activities that go on in a language classroom, the subject of the remainder of Chapter 9.

The conclusions on the need to vary tasks in testing, in order to cover CLA fully, apply at least as much to learning activities. These too should be scrutinised, and manipulated where necessary, e.g. through role-play. However, the most important contribution of this study to language learning is probably what has emerged about smallword use. I trust that anyone involved in teaching, having read this book, will appreciate the significance of this body of language. If this is the case, then the problem will largely solve itself. Part of the reason for the neglect of smallwords in the classroom is the status they have traditionally (not) enjoyed. Few would dispute the necessity of devoting time to the most everyday vocabulary, the commonest grammatical structures and bodies of little words, such as articles and pronouns. Yet words as common as *well* and *right* (and things like that) have been ignored, and even stigmatised. Teachers who have overcome the hurdle of accepting smallwords are in a position to raise their learners' awareness of them and to build learning activities around them. Smallwords exist, after all, everywhere that language is spoken. They are captured in their hordes on video, and are abundantly visible in language corpora. And more enlightened books of classroom activities have begun to cater for them directly. 'Implications for teaching and learning' pages 237 – 239, presents some ideas and references for putting smallwords on the classroom agenda.

A small word in conclusion

This study began with a test of speaking that needed to be validated. The process was laborious at times, but I believe it was worth it. Not only did I find out more or less what I needed to know about my own test, and how I might improve it, but I made many discoveries en route. I found out how to actually go about testing a test. I gained a clearer idea of what fluency is and what makes it happen. Not only did I see clear evidence that smallwords do make the difference between more and less fluent speakers, but I also understood why, by discovering what it is that smallwords actually do. This discovery was the one that struck me most, for it lies at the heart of understanding how to make a real difference to the people who count most: language learners.

Glossary of terms

backchannels backchannels are described in Stenström (1994) as making interaction possible 'without proper turn-taking, namely in cases where there is a (temporarily) dominant speaker and the other party's contribution is reduced to so-called "backchannels" (realised by items *like m, yes, oh, I see, really*) as a sign of attention' (1994: 1)

band scale a set of scaled descriptions of performance at different levels – or bands – of ability

channel the actual way our speaking 'travels', e.g. face-to-face, by telephone

CLA communicative language ability

cognitive environment (in relevance theory) all the facts and assumptions that are manifest to a person, i.e. that s/he is potentially, although not necessarily consciously, 'aware of' and capable of conceptualising.

communication strategies ways of compensating for 'gaps' in language knowledge, e.g. by paraphrasing, or using a more general term

Communicative language ability the ability to use language to communicate, normally described as a set of components or constructs of ability

communicative competence the underlying language knowledge of a person combined with their ability to use it in real communication

conditions the contextual conditions that affect our speech, i.e. purpose, interlocutor, setting and channel

constructs definitions of abilities that are sufficiently tangible that we can make hypotheses about how they relate to actual behaviour, or to each other (adapted from Bachman 1990: 255)

descriptors the descriptions, in a band scale, of performance or ability at the various levels

diagnostic testing testing whose main purpose is to identify testees' strengths and weaknesses

discourse markers smallwords 'used to organise and hold the turn and to mark boundaries in the discourse' (from Stenström 1994: 63)

disruptive pause a pause which interrupts the natural flow of speech, since it does not occur at the boundary of some kind of unit of information

domain the area of language use for which CLA is defined, in terms of medium, topics, conditions and functions

epistemic modals expressions such as *I think*, which are used to state that we believe something to be the case, rather than to assert that it is the case

face a person's 'face' is what we give regard to when we try to make them feel 'good', or at least try to avoid making them seem insignificant, wrong, mean, stupid, etc. This may be done through the use of hedges

fluency
i) *(preliminary working definition)*
the ability to make decisions rapidly, implement them smoothly, and adjust our conversation as unexpected problems appear in our path (adapted from Bygate 1987: 3)
ii) *(definition arrived at in the course of the study)*
the ability to contribute what a listener, proficient in the language, would normally perceive as coherent speech, which can be understood without undue strain, and is carried out at a comfortable pace, not being disjointed, or disrupted by excessive hesitation

formulaic expressions expressions that are produced as ready-made chunks of language rather than composed by the speaker

functions the things we actually 'do' when we speak – e.g. asking, apologising

G-theory G (generalisability)-theory provides a way of identifying and estimating the significance of the different factors that affect test score

hedge softening the force of what we are saying is 'hedging', and this is done with hedges such as *I think*

initial context (in relevance theory) the context created as a result of what has just been said, often interacting with, for example, shared knowledge or the physical environment

interaction any conversation or other speaking situation where people are talking to each other

interactional signals smallwords used to involve or acknowledge the interlocutor or what s/he has said (adapted from Stenström 1994: 63)

interlocutor a person taking part in an interaction

macro-functions the broad categories of functions, e.g. imparting and seeking information, or socialising

macroskills often referred to as 'the four skills': listening, speaking, reading and writing

medium the 'basic type' of communication i.e. speaking or writing

microlinguistic ability the ability to access and use with some degree of correctness the essential systems of language at the level of the sentence/utterance and below – vocabulary, morphology, syntax and phonology

operationalisation the definition of a construct (or component) of CLA in terms of actual language behaviour

ostensive-inferential communication (in relevance theory) verbal communication is 'ostensive' because the speaker's utterance makes plain what s/he is intending to do in the act of communication, and 'inferential' because the hearer is expected to infer the speaker's meaning from what is said

performance profile form a form containing questions about different aspects of test performance, which, when filled in, provides a profile of the testee's performance

positive cognitive effect (in relevance theory) a change (i.e. in line with the speaker's intentions) brought about, by what is communicated, in the hearer's cognitive environment

pragmatic ability the ability to use and interpret language in the way that it is typically used and interpreted by the society and in the particular situation in which the communication is taking place

routines predictable patterns in speech, e.g. opening or closing telephone conversations, or buying things

score here the term is used as by Messick (1995) 'generically in its broadest sense to mean any coding or summarisation of observed consistencies or performance regularities on a test, questionnaire, observation procedure, or other assessment devices such as work samples, portfolios and realistic problem simulations' (1995: 741)

skill the particular things people have to be able to do, specifically when communicating through speaking

smallwords the small words and expressions that help to keep our speech flowing, yet do not contribute essentially to the message itself

strategic ability the ability to use devices to keep conversation going in the face of difficulty and to check for, explain and tackle potential problems in communication

T-unit a unit of speech consisting of a main clause and any subordinate clauses attached to it or embedded in it

temporal variables variables that are related to the speed of speaking, such as actual speech rate, pausing and anything that breaks up or slows down the flow of speech

textual ability the ability to make a text 'coherent with respect to itself', involving cohesion as the expression of semantic relations and the use of markers and routines that build structure into conversation as well as the organisational ability to structure information

TLU target language use

utterance a fragment of speech that actually 'says' something, or 'the physical realisation of a sentence' (Wales 1989: 471)

vague language the words and phrases that express vagueness or lack of commitment, e.g. *sort of, I think*

washback effect the effect a test has on what goes on in the classroom, or on any other education-related process

References

Aijmer, K. (1987) 'Oh and Ah in English Conversation', in W. Meijs (ed.) *Corpus Linguistics and Beyond.* Amsterdam: Rodopi, 61 – 86.

Aijmer, K. (1996) *Conversational Routines in English*. London: Longman.

Alderson, J. C. (1991) 'Bands and Scores'. In J. C. Alderson and B. North (eds.) *Language Testing in the 1990s.* London: Macmillan, 71 – 86.

Alderson, J. C., Clapham, C. and D. Wall. (1995) *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

Alderson, J. C. and L. Hamp-Lyons. (1996) TOEFL *Preparation Courses: a Study of Washback*. Language Testing 13(3) 280 – 297.

Alderson, J. C. and B. North (eds.) (1991) *Language Testing in the 1990s.* London: Macmillan.

Andersen, G. (1998) 'The Pragmatic Marker *Like* from a Relevance-Theoretic Perspective'. in A. Jucker and Y. Ziv (eds.) *Discourse Markers: Descriptions and Theory.* Amsterdam: John Benjamins, 147 – 170.

Anderson, J.R. (1983) *The Architecture of Cognition*. Cambridge, M A: Harvard University Press.

Atkinson, J. M. and J. Heritage (eds.) (1984) *Structures of Social Action*. Cambridge: Cambridge University Press.

Bachman, L. F. (1990) *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Bachman, L. F. and A. S. Palmer (1996) *Language Testing in Practice*. Oxford: Oxford University Press.

Bailey, K. M. (1996) 'Working for Washback: A Review of the Washback Concept in Language Testing'. *Language Testing*, 13(3) 257 – 279.

Baker, R. (1997) *Classical Test Theory and Item Response Theory in Test Analysis* (Special Report No 2, Language Testing Update). University of Lancaster: Centre for Research in Language Education.

Bialystok, E. (1990) *Communication Strategies.* Oxford: Blackwell.

BrainPower Inc. (1986) *Statview User Manual*. Calabasas: Abacus Concepts Inc.

Brown, G. and G. Yule (1983a) *Teaching the Spoken Language*. Cambridge: Cambridge University Press.

Brown, G. and G. Yule (1983b) *Discourse Analysis*. Cambridge: Cambridge University Press.

Brown, P. and S. Levinson (1987) *Politeness.* Cambridge: Cambridge University Press.

Bygate, M. (1987) *Speaking*. Oxford: Oxford University Press.

Canale, M. (1983) 'From Communicative Competence to Language Pedagogy'. In J. C. Richards and R. W. Schmidt (eds.) *Language and Communication.* London: Longman, 2 – 27.

Canale, M. and M. Swain (1980) Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing. *Applied Linguistics,* 1(1) 1 – 47.

Channell, J. (1994) *Vague Language.* Oxford: Oxford University Press.

Cheshire, J. (1998) Linguistic Variation and Social Function. In Coates, J. (ed.) *Language and Gender: A Reader.* Oxford: Blackwell, 29 – 41.

Child, D. (1990) *The Essentials of Factor Analysis.* London: Cassell Education Limited.

Chomsky, N. (1965) *Aspects of the Theory of Syntax*. Boston: MIT Press.

Clapham, C. (1988) 'Concurrent Validity'. In *English Language Testing Service. Validation Report (ii).* London and Cambridge: The British Council and Cambridge Local Examinations Syndicate, 49 – 53.

Clark, H. H. (1996) *Using Language.* Cambridge: Cambridge University Press.

Coates, J. (1986) *Women, Men and Language*. London: Longman.

Coates, J. (1998) *Language and Gender: A Reader.* Oxford: Blackwell.

Cook, G. (1997) *Language Play, Language Learning*. ELT Journal 51(3) 224 – 231.

Council of Europe (1988) *Project no. 12: Evaluation and Testing in the Learning and Teaching of Languages for Communication*. Strasbourg: Council of Europe.

Council of Europe (1992) *Transparency and Coherence in Language Learning in Europe; Objectives, Assessment and Certification.* Strasbourg: Council of Europe.

Council of Europe (1996) *Common European Framework of Reference for Language Learning and Teaching.* Strasbourg: Council of Europe.

Council of Europe (2001) *Common European Framework for Languages: Learning, Teaching, Assessment.* Cambridge: Cambridge University Press.

Criper, C. and A. Davies (1988) *English Language Testing Service. Validation Report (i).* London and Cambridge: The British Council and Cambridge Local Examinations Syndicate.

Crystal, D (1991) *A Dictionary of Linguistics and Phonetics.* Oxford: Blackwell.

Cumming, A. and R. Berwick (eds.) (1996) *Validation in Language Testing*. Clevedon: Multilingual Matters.

Cumming. A. (1996) 'Introduction: The Concept of Validation in Language Testing'. In Cummins, A. and R. Berwick (eds.) *Validation in Language Testing*. Clevedon: Multilingual Matters, 1 – 14.

Dagut, M.B. (1977) Incongruencies in Lexical 'Gridding'. *IRAL* 15(3), 221 – 229.

Davies, A. (1990) *Principles of Language Testing.* Oxford: Blackwell.

De Bot, K. (1992) 'A Bilingual Production Model. Levelt's 'Speaking' Model Adapted'. *Applied Linguistics* 13(1) 1 – 24.

De Cock, S., S. Granger, G. Leech and T. McEnery (1997) 'An Automated Approach to the Phrasicon of EFL Learners'. In S. Granger (ed.) *Learner English on Computer*. London: Longman, 67 – 79.

De Jong, J. H. A. L. (1991) *Defining a Variable Foreign Language Ability*. (Doctoral Dissertation) The Hague: CIP– Gegevens Koninkluke Biblioteek.

Dechert, H. W., Möhle D. and M. Raupach (eds.) (1984) *Second Language Productions*. Tübingen: Narr.

Dörnyei, Z. and S. Thurrell (1992) *Conversation and Dialogues in Action.* London: Prentice Hall.

Douglas, D. (1994) *Quality and Quantity in Speaking Test Performance*. Language Testing, 11(2) 125 – 144.

Esser, U. (1996) *Oral Language Testing: The Concept of Fluency Revisited*. Lancaster University: Unpublished MA Dissertation.

Færch, C. and G. Kasper (1983) *Strategies in Interlanguage Communication*. London: Longman.

Freed, B. (1995) (ed.) *Second Language Acquisition in a Study Abroad Context*. Amsterdam: John Benjamins.

Freed, B. (1995) 'What Makes Us Think that Students who Study Abroad Become Fluent?'. In B. Freed (ed.) *Second Language Acquisition in a Study Abroad Context*. Amsterdam: John Benjamins, 123 – 148.

Fulcher, G. (1993) *The Construction and Validation of Rating Scales for Oral Tests in English as a Foreign Language*. (PhD Thesis). Lancaster: University of Lancaster.

Fulcher, G. (1996) 'Does Thick Description Lead to Smart Tests? A Data-Based Approach to Rating Scale Construction'. *Language Testing, 13(2*) 208 – 238.

Gass, S. and L. Selinker (eds.) (1983) *Language Transfer in Language Learning.* Rowley, MA: Newbury House.

Granger, S. (ed.) (1997) *Learner English on Computer*. London: Longman.

Granger, S. (1997) 'The Computer Learner Corpus: A Versatile New Source of Data for SLA Research'. In S. Granger (ed.) *Learner English on the Computer*. London: Longman, 3 – 18.

Granger, S., J. Hung and S. Petch-Tyson (eds.) (2002) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins.

Greenbaum, S., G. Leech and J. Svartvik (eds.) (1980) *Studies in English Linguistics for Randolph Quirk*. London: Longman.

Halliday, M. A. K. (1994) *Functional Grammar*. London: Edward Arnold.

Halliday, M. A. K. and R. Hasan (1976) *Cohesion in English*. London: Longman.

Hamp-Lyons, L. (1997) 'Washback, Impact and Validity: Ethical Concerns'. *Language Testing*, 14(3) 295 – 303.

Harley, B., P., Allen, J. Cummins and M. Swain (eds.) (1990) *The Development of Second Language Proficiency.* Cambridge: Cambridge University Press.

Hasselgren, A. (1993) Lexical Teddy Bears and Advanced Learners: A Study into the Way Norwegian Students Cope with English Vocabulary. *International Journal of Applied Linguistics*, 237 – 260.

Hedge, T. (1993) Key Concepts in ELT. In *ELT Journal*, 47(3) 275 – 277.

Henning, G. (1987) *A Guide to Language Testing: Development, Evaluation, Research.* Rowley: Newbury House.

Heritage, J. (1984) A Change-of-State Token and Aspects of its Sequential Placement. In J. M. Atkinson, and J. Heritage (eds.) (1984) *Structures of Social Action*. Cambridge: Cambridge University Press, 299 – 344.

Holmes, J. (1984) Modifying Illocutionary Force. *Journal of Pragmatics 8,* 345 – 365.

Holmes, J. (1988) *Sort of* in New Zealand Women's and Men's Speech. *Studia Linguisitica, 42(2)* 85 – 117.

Holmes, J. (1995) *Women, Men and Politeness.* London: Longman.

Hughes, A. (1989) *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Hughes, A., D. Porter and C. Weir (eds.) (1988) *English Language Testing Service, Research Report 1 (ii)*. Cambridge: The British Council and the University of Cambridge Examinations Syndicate.

Huhta, A., V. Kohonen, L. Kurki-Suonio and S. Luoma (1997) *Current Developments and Alternatives in Language Assessment. Proceedings of LTRC 96.* Jyväskylä: University of Jyväskylä.

Huhta, A., K. Sajavaara and S. Takala (eds.) (1993) *Language Testing: New Openings*. Jyväskylä: Institute for Educational Research, University of Jyväskylä.

Hymes, D. (1972) On Communicative Competence. In J. B. Pride and J. Holmes (eds.) *Sociolinguistics*. Harmondsworth: Penguin, 269 – 293.

James, D. (1978) The Use of *oh, ah, say* and *well* in Relation to a Number of Grammatical Phenomena. In *Papers in Linguistics ll: 3 – 4,* 517 – 535.

Jucker, A and Y. Ziv (eds.) (1998) *Discourse Markers: Descriptions and Theory.* Amsterdam: John Benjamins.

Jucker, A. (1993) The Discourse Marker *well*: a Relevance Theory Account. *Journal of Pragmatics 19,* 435 – 452.

Kellerman, E. (1983) Now You See it, Now You Don't. In S. Gass and L. Selinker (eds.) *Language Transfer in Language Learning.* Rowley, MA: Newbury House, 112 – 134.

Kellerman, E., T. Bongaerts and N. Poulisse (1987) Strategy and System. In L2 Referential Communication, in R. Ellis (ed.) *Second Language Acquisition in Context.* London: Prentice Hall, 110 – 112.

Koponen, M. (1995) Let Your Language and Speech Flow! Is there a Case for the Construct of Fluency in Perception of Oral Performance? *Unpublished Proceedings from Language Testing Forum, Newcastle University, November 24 – 26 1995.*

Krashen, S. (1982) *Principles and Practice in Second Language Acquisition.* Oxford: Pergamon.

Labov, W. (1972) *Sociolinguistic Patterns*. Philadelphia: Pennsylvania University Press.

Lakoff, G. (1982). 'Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts'. *Papers from the Eighth Regional Meeting, Chicago Linguistics Society*, 183 – 228.

Lakoff. R. (1973) Questionable Answers and Answerable questions. *Issues in Linguistics. Papers in Honor of Henry and Renee Kahane.* Urbana, Il: University of Illinois Press, 453 – 467.

Lee, D. (1987) The Semantics of Just. *Journal of Pragmatics* 11, 377 – 398.

Lennon, P. (1990) Investigating Fluency in EFL: A Quantitative Approach. *Language Learning*, 40(3) 387 – 417.

Levelt, W. J. M. (1983) 'Monitoring and Self-Repair in Speech'. *Cognition* 14, 41 – 104.

Levelt, W. J. M. (1989) *Speaking: – from Intention to Articulation*. Cambridge, Mass: MIT Press.

Levinson, S. C. (1983) *Pragmatics*. Cambridge: Cambridge University Press.

Lightbown, P. M. and N. Spada (1993) *How Languages are Learnt.* Oxford: Oxford University Press.

Little, D. and R. Perclová (2001) *The European Language Portfolio: A Guide for Teachers and Teacher Trainers*. Strasbourg: Council of Europe Publications.

McNamara, T. (1996) *Measuring Second Language Performance*. New York: Addison-Wesley Longman.

Meijs, W. (ed.) (1987) *Corpus Linguistics and Beyond,* Amsterdam: Rodopi.

Messick, S. (1989) Validity. In Linn, R. L. (ed.) *Educational Measurement* (3rd Edition). New York: Macmillan, 13 – 103.

Messick, S. (1994) 'The Interplay of Evidence and Consequences in the Validation of Performance Assessments'. Educational Researcher 23(2), 13 – 23.

Messick, S. (1995) Validity of Psychological Assessment. *American Psychologist*, 50 (9), 741 – 749.

Messick, S. (1996) Validity and Washback in Language Testing. *Language Testing* 13(3) 241 – 256.

Mey, J. L. (1993) *Pragmatics – an Introduction.* Oxford: Blackwell.

Milroy, L. (1987) *Observing and Analysing Natural Language*. Oxford: Blackwell.

Nattinger J. R. and J. S. DeCarrico (1992) *Lexical Phrases and Language Teaching.* Oxford: Oxford University Press.

Nikula, T. (1996) *Pragmatic Force Modifiers.* Jyväskylä: University of Jyväskylä.

Nolasco, R. and L. Arthur (1987) *Conversation*. Oxford: Oxford University Press.

North, B. (1992) A European Language Portfolio: Options for Scales for Proficiency. In Council of Europe *Transparency and Coherence in Language Learning in Europe; Objectives, Assessment and Certification*. Strasbourg: Council of Europe, 158 – 174.

North, B. (1997) The Development of a Common Framework for the Description of Language Learning, Teaching and Assessment. In A. Huhta, V. Kohonen, L. Kurki-Suonio and S. Luoma (eds.) *Current Developments and Alternatives in Language Assessment. Proceedings of LTRC 96.* Jyväskylä: University of Jyväskylä, 423 – 447.

Norwegian Ministry of Education and Research (1987) *M87 – The Curriculum Guidelines for Compulsory Education in Norway.* 1987. Oslo. The Ministry of

Education and Research and Aschehoug.

Odlin, T. (1989) *Language Transfer.* Cambridge: Cambridge University Press.

Oskarrson, M. (1988) Self-Assessment of Communicative Proficiency. In Council of Europe, Project no. 12: *Evaluation and Testing in the Learning and Teaching of Languages for Communication*. Strasbourg: Council of Europe, 46 – 58.

Overstreet, M and G. Yule (1997) On Being Inexplicit and Stuff in Contemporary American English. *Journal of English Linguistics* 25(3), 250 – 258.

Pawley, A. and F. H. Syder (1983) Two Puzzles for Linguistic Theory: Nativelike Selection and Nativelike Fluency. In J. C. Richards and R. W. Schmidt (eds.) *Language and Communication.* London: Longman, 191 – 227.

Peters, A. (1983) *The Units of Language Acquisition*. Cambridge: Cambridge University Press.

Pienemann, M. and M. Johnston (1987) Factors Influencing the Development of Fluency. In D. Nunan (ed.) *Applying Second Language Research*. Adelaide: National Curriculum Resource Centre, 45 – 141.

Pride, J. B. and J. Holmes (eds.) (1972) *Sociolinguistics*. Harmondsworth: Penguin.

Raupach, M. (1984) Formulæ in Second Language Speech Production. In H. W. Dechert, D. Möhle and M. Raupach (eds.) *Second Language Productions*. Tübingen: Narr, 114 – 137.

Richards, J. C. and R. W. Schmidt (eds.) (1983) *Language and Communication.* London: Longman.

Ringbom, H. (1987) *The Role of the First Language in Foreign Language Learning.* Clevedon: Multilingual Matters.

Savignon, S. J. (1983) *Texts and Contexts in Second Language Learning*. Reading, MA: Addison-Wesley.

Schachter, J. (1990) Communicative Competence Revisited. In B. Harley, P. Allen, J. Cummins and M. Swain (eds.) *The Development of Second Language Proficiency.* Cambridge: Cambridge University Press, 39 – 49.

Schiffrin; D. (1987) *Discourse Markers.* Cambridge: Cambridge University Press.

Schourup, L. (1985) *Common Discourse Particles in English Conversation.* New York: Garland Publication.

Searle, J. R. (1969) *Speech Acts.* Cambridge: Cambridge University Press.

Seda, I. and S. Abrahamson (1990) English Writing Development of Young, Linguistically-Different Learners. *Early Childhood Research Quarterly* 5, 379 – 391.

Sharwood Smith. M. (1994) *Second Language Learning*. London: Longman.

Shohamy, E. (1993) The Exercising of Power and Control in the Rhetorics of Testing. In A. Huhta, K. Sajavaara and S. Takala (eds.) *Language Testing: New Openings*. Jyväskylä. Institute for Educational Research, University of Jyväskylä, 23 – 38.

Shohamy, E. (1994) The Validity of Direct Versus Semi-Direct Oral Tests. *Language Testing* 11(2), 99 – 123.

Shohamy, E., R. Reves and Y. Bejarano (1986) *Introducing a New Comprehensive*

Test of Oral Proficiency. ELT Journal, 40(3) 212 – 220.

Sinclair, J. (1991) *Corpus, Collocation, Concordance*. Oxford: Oxford University Press.

Skehan, P. (1991) Progress in Language Testing in the 1990s. In J. C. Alderson and B. North (eds.). *Language Testing in the 1990s.* London: Macmillan, 3 – 21.

Sperber, D. and D. Wilson (1995) *Relevance*. Oxford. Blackwell.

Stenström, A-B. (1984) *Questions and Responses in English Conversation*. Lund. Lund University Press.

Stenström, A-B. (1990) Lexical Items Peculiar to Spoken Discourse. In J. Svartvik. *The London-Lund Corpus of Spoken English.* Lund: Lund University Press, 137 – 175.

Stenström, A-B. (1994) *An Introduction to Spoken Interaction.* London: Longman.

Svartvik J. (1990) The *London-Lund Corpus of Spoken English.* Lund: Lund University Press.

Svartvik, J. (1980) Well in Conversation. In S. Greenbaum, G. Leech and J. Svartvik (eds.) *Studies in English Linguistics for Randolph Quirk.* London: Longman, 162 – 177.

Tarone, E. and G. Yule (1989) *Focus on the Language Learner.* Oxford: Oxford University Press.

Towell, R., R. Hawkins, and N. Bazergui (1996) The Development of Fluency in Advanced Learners of French. *Applied Linguistics* 17(1) 84 – 119.

Van Ek, J. A. and J. L. M. Trim (1993) *Threshold Level 1990.* Strasbourg: Council of Europe.

Wales, K. (1989) *A Dictionary of Stylistics*. London: Longman.

Wall, D., C. Clapham, and J. C. Alderson (1991) Validating Tests in Difficult Circumstances. In C. Alderson and B. North (eds.) *Language Testing in the 1990s*. London: Macmillan, 209 – 225.

Wall, D., C. Clapham, and J. C. Alderson (1994) Evaluating a Placement Test. *Language Testing*, 11(3) 321 – 343.

Weir, C. (1988) Construct Validity. In A. Hughes, D. Porter and C. Weir (eds.) *English Language Testing Service, Research Report 1 (ii)*. London and Cambridge: The British Council and the University of Cambridge Local Examinations Syndicate, 10 – 14.

Weir, C. (1993) *Understanding and Developing Language Tests*. London: Prentice-Hall.

Wong-Fillmore. L. (1976) *The Second Time Around: Cognitive and Social Strategies in Second Language Acquisition.* Unpublished Doctoral Dissertation. Ann Arbor, MI: Stanford University.

Yule, G. (1996) *Pragmatics*. Oxford: Oxford University Press.

Appendices

Appendix A

Version of EVA speaking test

Task 1 pupil A

All the lonely people

You each have a set of pictures. Your pictures make the first half of a story, telling about 24 hours in Paul's life. Your friend has the second half of the story.

Look at the pictures.

Listen to **B** describing one of his/her pictures

Now you **describe** your biggest picture. Give as much detail as you can.

Now look at the pictures and then tell your part of the story. Tell it as a **story** of 24 hours in Paul's life.

What sort of feelings do you think Paul has in pictures 2, 4 and 5?

What do you think happens next?

Listen to **B** telling what happens.

What do you think?.....
(Take turns answering these questions. Agree or disagree with each other).

Is there a 'message' in the story?

What might the title be?

Finish this sentence: The worst thing about loneliness is...

Have you ever felt lonely like Paul, or do you know someone who might?

What can make a 14-15 year old feel 'alone'?


1
PAUL COLLECTS STAMPS...
2
3
0,50
HA HA HA HA HA
PAUL
4
5

Task 2 pupil A

1 Working in the garden

You sometimes do jobs for an old lady, in her garden, for pocket money. This weekend you are going to be away. Your friend says she/he will do the jobs instead of you. You have to tell him/her what to do.

Look at the 4 pictures and decide on the job for each picture.

Tell your friend

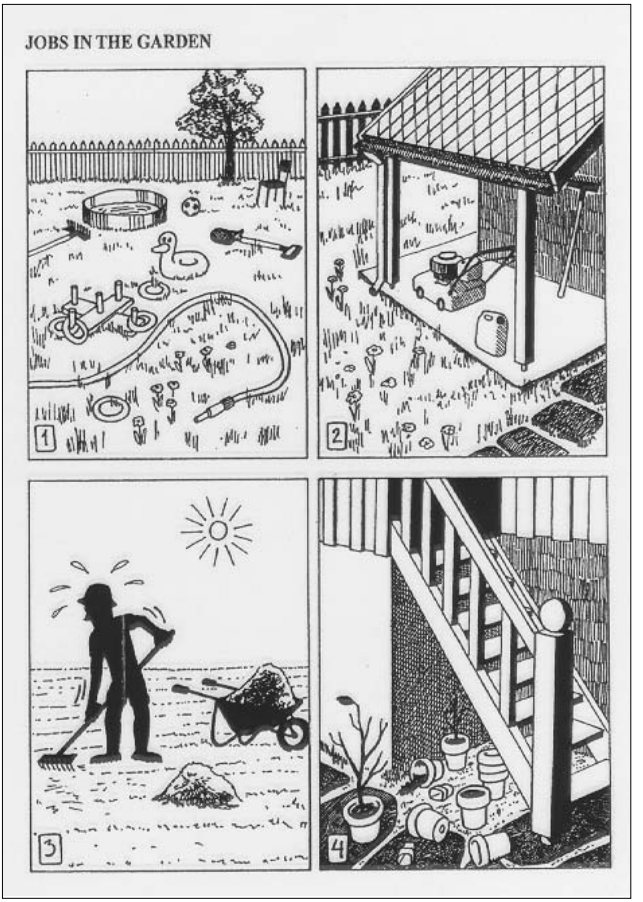
- what the 4 jobs are
- exactly how to do the jobs
- anything s/he needs to know (e.g. where things are, or should be put)

Then check that s/he really knows what to do. **Ask her/him to repeat the main points.** (It's very important that the jobs are done properly!)
(Help her/him to get it right)

2 Fetching things from the bedroom

Your friend needs some things from his/her bedroom. You are going to fetch them. You can see them in the pictures below. Your friend will tell you exactly where they are. **Listen carefully. Ask** if you don't understand.

At the end, check that you have understood and remembered. **Try to tell your partner where the things are.**


JOBS IN THE GARDEN
1
2
3
4

Task 3 part 2

Phoning about a holiday job

Pupil B
Read aloud what is in the box.

It is summer. You are staying with relatives in England for two months until the middle of August, and would like to earn some money. You see this advert in the local newspaper.

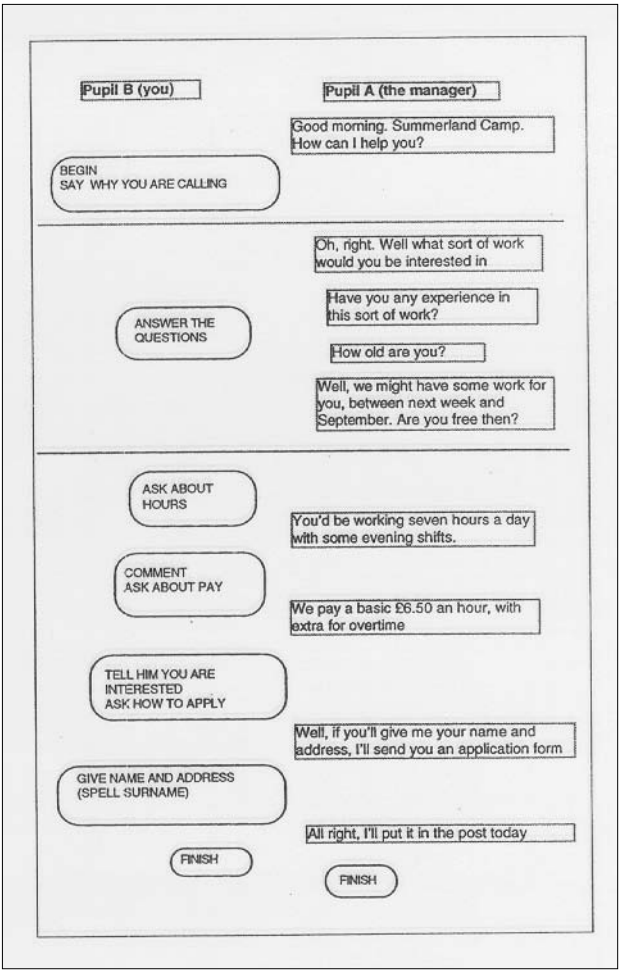
> *Want a holiday job*
> Holiday Camp needs assistans for summer season. If you are interested in serving in our restaurants, organising sports and games for children, or cleaning rooms, please call us on 74.612395.

This could be just what you are looking for! You phone to ask for more details. Talk to the manager. The 'boxes' tell you what to do. You can put in extra comments and phrases to make the conversation 'flow'.

(Decide which type of work you would be most interested in.)

your address: 3 Cliff Top Lane
Whitby

Pupil A
You are the manager at the holiday camp. You answer the phone and talk to pupil B. Your part is written down. You have to read it. You start by answering the phone.


Pupil B (you)
Pupil A (the manager)
Good morning. Summerland Camp. How can I help you?
BEGIN
SAY WHY YOU ARE CALLING
Oh, right. Well what sort of work would you be interested in
Have you any experience in this sort of work?
ANSWER THE QUESTIONS
How old are you?
Well, we might have some work for you, between next week and September. Are you free then?
ASK ABOUT HOURS
You'd be working seven hours a day with some evening shifts.
COMMENT
ASK ABOUT PAY
We pay a basic £6.50 an hour, with extra for overtime
TELL HIM YOU ARE INTERESTED
ASK HOW TO APPLY
Well, if you'll give me your name and address, I'll send you an application form
GIVE NAME AND ADDRESS (SPELL SURNAME)
All right, I'll put it in the post today
FINISH
FINISH

Appendix B

Performance profile

EVA test of spoken English **Performance profile** Pupil A/B.....

This marking scheme is intended as a guide for the tester to be consulted when deciding the pupil's grade, and as a form of feedback to the pupil.

task 1 – picture and story/account of events

i) *the picture was presented*

clearly and independently, with detail and as a 'scene'	3 [ ]
with some detail, perhaps with some prompting	2 [ ]
only able to name features or simple happenings	1 [ ]

ii) *the story/account of events was put across*

independently and coherently, with some detail to add interest	3 [ ]
as essential facts, but with little detail or linking into a 'whole'	2 [ ]
barely at all	1 [ ]

iii) *feelings were expressed*

finely, e.g. 'relieved, worried'	3 [ ]
roughly but adequately on the whole, e.g. 'sad, happy'	2 [ ]
inadequately	1 [ ]

iv) *opinions were offered*

freely with clear and complete reasons	3 [ ]
simply, with some reasons, perhaps with encouragement	2 [ ]
only as brief 'yes/no' type opinions	1 [ ]

task 2 – instructions

i) *the instructions were given*

clearly, coherently and independently	3 [ ]
quite well, followable on the whole, maybe with some prompting	2 [ ]
only as main points in answer to questions from the tester	1 [ ]

ii) **the vocabulary was*

full and appropriate, with very few gaps	3 [ ]
simple but coped with on the whole, using English	2 [ ]
lacking considerably, with resorting to Norwegian or 'stopping'	1 [ ]

task 3 – the semi-role-play

a) In the reading,

i) **all the words were pronounced*

very well	3 [ ]
acceptably, allowing an occasional mispronunciation or problem	2 [ ]
badly, with many mispronunciations or problems	1 [ ]

ii) **the text*

'flowed' on the whole	3 [ ]
was quite broken up	2 [ ]
was very broken up	1 [ ]

iii) **the intonation*

was good and supported the message on the whole	3 [ ]
was not very good, but did not interfere with the message	2 [ ]
was poor, and interfered with the message	1 [ ]

b) In the conversation, was the pupil able to

i) **use the appropriate style and degree of politeness necessary to the situation*
(e.g. using 'please' with requests, thanking for favours, 'would you ...' etc.)

yes	3 [ ]
to some extent	2 [ ]
no	1 [ ]

ii) **sound interested and friendly*

yes	3 [ ]
fairly	2 [ ]
no	1 [ ]

iii) **use expressions particular to the situation and tackle special conventions like giving phone numbers, dates, spelling, etc.*

yes	3 [ ]
to some extent	2 [ ]
no	1 [ ]

iv) *put the whole 'message' across*

completely	3 [ ]
partly	2 [ ]
barely	1 [ ]

In the test generally

Pupil A/B.....

i) *(strategies) *when difficulties in communication arose, did the pupil make an independent attempt to overcome these, in English?*

virtually always	3 [ ]
sometimes	2 [ ]
rarely	1 [ ]

ii) **was the pupil able to take the initiative?*

yes, frequently	3 [ ]
yes, at times	2 [ ]
no	1 [ ]

iii) **was the pupil able to keep going without a lot of prompts?*

yes, frequently	3 [ ]
on the whole, but needed frequent encouragement	2 [ ]
no, was only able to respond to questions	1 [ ]

iv) **The pupil's contributions to the conversation were:*

very full and significant	3 [ ]
hesitant but significant	2 [ ]
barely significant	1 [ ]

v) **the pronunciation was*

very good, with no sounds that could be misinterpreted	3 [ ]
good enough to get the message across	2 [ ]
difficult to understand (or would be for a non-Norwegian!)	1 [ ]

vi) **the intonation*

was good and supported the message on the whole	3 [ ]
was not very good, but did not interfere with the message	2 [ ]
was poor, and interfered with the message	1 [ ]

vii) **did the tone indicate friendliness, politeness and interest?*

very much so	3 [ ]
to some extent	2 [ ]
no	1 [ ]

viii) *(correctness) *the language structures and vocabulary were*

appropriate and fully understandable, without many errors	3 [ ]
on the whole understandable, despite errors and gaps	2 [ ]
so full of errors and gaps that the message was not easily understood	1 [ ]

ix) **the language choices were*

adventurous, independent and idiomatic	3 [ ]
independent and idiomatic at times	2 [ ]
very dependent on the 'given' material/input	1 [ ]

Appendix C

Band scales with guidelines

Band scales for spoken English

MESSAGE AND FLUENCY

5 – 6

This level is characterised by a good and independent performance on all tasks tested. The pupil should take the initiative and willingly supply original, detailed contributions, which are linked logically into cohesive (*sammenhengende*) 'wholes'. S/he should be able to 'keep going' with the minimum of help. Speech 'will be 'flowing' with little hesitation.

For the final awarding of '5' or '6', see the section on *pronunciation and intonation.*

3 – 4

At this level the pupil should manage an adequate performance on most of the tasks, with some simple linking of ideas.

At level 3, the pupil may be very hesitant and need a good deal of help to keep going. S/he should manage 'the essentials' of tasks that involve concrete ideas, such as information, actions and physical features in pictures.

At level 4, the pupil will sometimes be able to keep going quite well without help. S/he will from time to time take the initiative and contribute more than the 'essentials' of a task. Both concrete and more abstract ideas, such as opinions, reasons and feelings, will be tackled fairly well.

1 – 2

What the pupil contributes at this level will largely be in response to questions from the tester, and answers will be very short and generally inadequate.

At level 2 however, the pupil should make at least a minimal response to most tasks.

LANGUAGE STRUCTURES AND VOCABULARY

5 – 6

At this level the language should be characterisable as idiomatic, varied, independent and with few errors. The vocabulary will have very few 'gaps'. The pupil should be able to use the appropriate style and degree of politeness. The pupil will rarely resort to using Norwegian.

For the final awarding of '5' or '6', see section on *pronunciation and intonation.*

3 – 4

At this level there will be errors and Norwegianisms, but the message should be understandable in most tasks. Ideas will be linked simply.

At level 4, the language should show some originality and independence from the given material. Norwegian will only be resorted to from time to time.

At level 3, the wording may not be very original, being very dependent on what is 'given' in the task. Norwegian words may often be used.

1 – 2

At this level, the language will be insufficient to cope with the more demanding tasks.

However, at level 2, despite 'gaps', many errors and frequent reliance on Norwegian terms, the language should be sufficient to provide simple short responses to most tasks.

Setting grades on the oral test
After listening to the tape and filling in the performance profile for each pupil, the next task is to decide on a grade – from 1 to 6 (corresponding to Lg, Ng, G-, G+, M and S in the school system). The criteria scales (shown at the end of this section) are used for this purpose. The following advice is offered:

Using the performance profile as a guide:
1 Try to match the performance with a grade on each of the two scales: *message and fluency* and *language structures and vocabulary* (no performance will match perfectly – try to find the nearest description). It is a good idea to begin by deciding which of the three broad 'bands' the performance best fits into, and then seeing whether the performance would be 'weak' or 'strong' within this band – this will indicate the approximate grade.
2 Find an 'overall grade' that best reflects the grades on the two scales.
3 Use *pronunciation/intonation* as a final adjuster (see later below) in order to arrive at a final grade.

In placing a pupil at an overall grade, the general principle employed is that the teacher should aim at choosing the grade that best reflects the placings on both scales. However, when this decision is difficult to make, the scale of message and fluency should be weighted more heavily than the scale related to language.

There are two related reasons for this: firstly, in the speaking situation, the success of communication is dependent on numerous factors, of which linguistic 'correctness' is only one. Being able to take the initiative and 'keep going', showing politeness and interest, and knowing how to cope when language limitations arise are crucially important in speaking.

The second reason for rating 'communicativeness' over 'correctness', like the first, lies in the nature of speaking itself. Speaking is done under time pressure, and with no means of 'crossing out' what was incorrect. Even native speaker speech is 'untidy' compared to writing. We make false starts and rarely produce a perfect 'sentence'.

Learners have the additional burden of coping with a second language grammar under pressure, and even very advanced foreign speakers of English make mistakes that they would never submit on paper.

It seems reasonable therefore that pupils who will normally be assessed most critically in their written English on formal skills should be judged more mildly for these in the oral test, and credited for other, more relevant skills. This will give a more balanced total picture of the pupils' overall ability. If we were to use the same criteria for judging spoken and written performance, the oral test would have little function!

The criteria scales are shown at the end of this section.

The place of intonation and pronunciation

Because there is little direct correlation between pronunciation and intonation on the one hand and general performance on the other, pronunciation and intonation will be assessed independently and used as adjusters in setting the final grade given to a pupil.

The following guidelines are given:

At level 6, both pronunciation and intonation must be 'very good', i.e. the pronunciation is such that no sounds could be misinterpreted, and the intonation both supports the message and indicates friendliness and interest.

At level 5, both pronunciation and intonation must be 'acceptable', i.e. they do not 'block' the communication to a significant extent.

At all other levels, pupils on the 'borderline' between two grades should be upgraded when pronunciation and intonation are acceptable, and downgraded when they are not acceptable.

Appendix D

Guidelines for scoring procedure

Guidelines: scoring procedure

In this section, the evaluation process is described and advice is given on how to reach an overall assessment of spoken ability.

The process of evaluation has three stages. The first stage entails getting a picture or profile of how the pupil performed on the different parts of the test; the next stage involves 'placing' the pupil's performance according to the criteria scales, shown later in this section, leading to the awarding of a grade. Finally, the grade is adjusted if necessary according to the criteria relating to pronunciation and intonation.

The evaluator first listens to the performance of both pupils (A and B), while filling in a performance profile scheme for each pupil. These schemes are presented in the Appendix.

The format of the scheme follows the test, and two or three questions per task are asked, relating to those aspects of speaking highlighted by the particular task. The evaluator has to choose between three alternative levels of performance for each of the questions. Crosses may also be placed 'in-between' levels. Additionally, at the end there is a list of questions relating to the test generally. These should be read first, and borne in mind during the test.

These schemes are not intended to be used as 'score sheets'. Teachers should not add up the points given – this would lead to a wrongly weighted score. The absolute 'values' a teacher chooses are perhaps not so important as the overall profile that emerges for each pupil. The teacher can then use this profile when placing the pupil according to the criteria scales. Additionally, the scheme provides an excellent record, which clearly shows the relative strengths and weaknesses of the pupil.

What is more, by working through the profile when listening to the performance, the evaluator's attention is drawn to all the different aspects of speech that need to be considered when assessing the pupil. Without this kind of guidance it is tempting to 'listen for' certain limited aspects of speaking.

Setting grades on the oral test

After listening to the tape and filling in the performance profile for each pupil, the next task is to decide on a grade – from 1 to 6 (corresponding to Lg, Ng, G-, G+, M and S in the school system). The criteria scales (shown at the end of this section) are used for this purpose. The following advice is offered:

Using the performance profile as a guide:

1 Try to match the performance with a grade on each of the two scales: *messagee and fluency* and *language structures and vocabulary* (no performance will match perfectly – try to find the nearest description). It is a good idea to begin by deciding which of the three broad 'bands' the performance best fits into, and then seeing whether the performance would be 'weak' or 'strong' within this band – this will indicate the approximate grade.

2 Find an 'overall grade' that best reflects the grades on the two scales.

3 Use *pronunciation/intonation* as a final adjuster (see later below) in order to arrive at a final grade.

In placing a pupil at an overall grade, the general principle employed is that the teacher should aim at choosing the grade that best reflects the placings on both scales. However, when this decision is difficult to make, the scale of message and fluency should be weighted more heavily than the scale related to language.

There are two related reasons for this: firstly, in the speaking situation, the success of communication is dependent on numerous factors of which linguistic 'correctness' is only one. Being able to take the initiative and 'keep going', showing politeness and interest, and knowing how to cope when language limitations arise are crucially important in speaking.

The second reason for rating 'communicativeness' over 'correctness', like the first, lies in the nature of speaking itself. Speaking is done under time pressure, and with no means of 'crossing out' what was incorrect. Even native-speaker speech is 'untidy' compared to writing. We make false starts and rarely produce a perfect 'sentence'. Learners have the additional burden of coping with a second-language grammar under pressure, and even very advanced foreign speakers of English make mistakes that they would never submit on paper.

It seems reasonable therefore that pupils, who will normally be assessed most heavily in their written English on formal skills should be judged more mildly for these in the oral test, and credited for other, more relevant, skills. This will give a more balanced total picture of the pupils' overall ability. If we were to use the same criteria for judging spoken and written performance, the oral test would have little function!

The criteria scales are shown at the end of this section.

The place of intonation and pronunciation

Because there is little direct correlation between pronunciation and intonation on the one hand and general performance on the other, pronunciation and intonation will be assessed independently and used as adjusters in setting the final grade given to a pupil.

The following guidelines are given:

At level 6, both pronunciation and intonation must be 'very good', i.e. the pronunciation is such that no sounds could be misinterpreted, and the intonation both supports the message and indicates friendliness and interest.

At level 5, both pronunciation and intonation must be 'acceptable', i.e. they do not 'block' the communication to a significant extent.

At all other levels, pupils on the 'borderline' between two grades should be upgraded when pronunciation and intonation are acceptable, and downgraded when they are not acceptable.

Appendix E

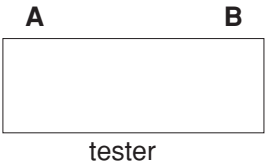
‘Script’ for test director

‘Script’ for test director – test version 3

General comments:
Try to stick roughly to this format, which is almost a ‘script’, to make sure all parts of tasks are covered, and to give the same ‘input’ to pupils.

Always name pupils clearly when starting them on any part of a task – this is very important for assessing from a recording!!

Tip: write down who is A and who is B. Place pupil A on the left (in relation to you) and B on the right, so they are lined up with their names on your paper.



The positioning shown here works well. Have the microphone as near as possible, without being intrusive. Don’t forget to record the test!!

Let the pupils have a quick look at the test book before they begin, and use a minute ‘warming up’, preferably in English.

Task 1

All the lonely people

In this task you are going to *describe, tell* and *discuss.*

You each have a set of pictures. **A**’s pictures make up the first half of a story, telling about 24 hours in Paul’s life. **B** has the second half of the story.

B, can you describe the big picture in the middle of your page? Give as much detail as you can.
Can you guess what the story is about?

Now **A**, can you please describe your biggest picture? Give as much detail as you can.
Now look at the pictures and then tell your part of the story. Tell it as a story of *24 hours in Paul’s life.*
What sort of feelings do you think Paul has in pictures 2, 4 and 5?
What do you think happens next?

Now **B**, can you tell what happens in the rest of the story? Tell it as the story of a *good day for Paul.*
How do you think Paul was feeling in pictures 8, 9 and 10?

What do you think?...

*(Take turns addressing questions to pupils **A** and **B**. After a pupil has given an answer, ask the other pupil if s/he agrees.)*
A: Is there a '*message*' in the story? (What do you think **B**?)
B: What might the title be? (What do you think **A**?)
A: Finish this sentence: *The worst thing about loneliness is ...*
(What do you think **B**?)
B: Have you ever felt lonely like Paul, or do you know someone who might have?
(What do you think **A**?)

(To both) What can make a 14–15-year-old feel 'alone'?

Task 2

In this task there are two parts. Both of them involve giving instructions. One of you has to give instructions. The other one has to listen carefully and check after that s/he knows what to do.

Working in the garden

A, You are going to start.

You sometimes do jobs for an old lady, in her garden, for pocket money. This weekend you are going to be away. Your friend says she/he will do the jobs instead of you. You have to tell him/her what to do.

Look at the 4 pictures and decide on the job for each picture.
Tell your friend:
- what the 4 jobs are
- exactly how to do the jobs
- anything s/he needs to know (e.g. where things are)

Then check that s/he really knows what to do. Ask her/him to repeat the main points. (It's very important that the jobs are done properly!)
(Help her/him to get it right)

B, Listen carefully and ask if there's anything you don't understand.
At the end, check that you know what to do.

Now it's **B**'s turn to give instructions.

Needing things from your bedroom
You are at school. You are going on a trip and you need some things from your bedroom. You aren't able to go home and fetch them. Your friend lives near you. S/he says s/he will get them. Tell what you want and where they are. The things are shown in the picture below and in the bedroom picture opposite.

Afterwards check that your friend knows where things are.
(Help her/him to get it right)

A, Listen carefully and ask if there's anything you don't understand.

At the end, check that you know what to do.

Task 3

In this task there are also two parts. Each one is a semi-role-play, where one pupil has to decide what to say, while the other mainly has to read his/her part.

The first part is called **At the hospital**

A, Will you please read aloud what is in the box?

> *You are on holiday in Ireland with your family. Your young sister was climbing a tree when she fell and hurt her arm. You think her arm may be broken. You take her straight to the casualty (akutthjelp) department at the local hospital. Your sister cannot speak English. You go to reception and talk to a nurse.*
>
> *Think of a name and a birth date for your sister.*

Look at the opposite page. Read what the nurse says and think about what you will say before you start the conversation. The 'boxes' tell you what to do. You can put in extra comments and phrases to make the conversation 'flow'.

B, You are the nurse. Talk to **A**. Your part is written down. You have to read it. You start the conversation.

The second part is called phoning about a holiday job

B, Will you read aloud what is in the box?

It is summer. You are staying with relatives in England for two months until the middle of August, and would like to earn some money. You see this advert in the local newspaper.

Want a holiday job?
Holiday Camp needs assistants for summer season. If you're interested in serving in our restaurants, organising sports and games for children, or cleaning rooms, please call us on 74 – 612395.

This could be just what you are looking for! You phone to ask for more details. Talk to the manager. The 'boxes' tell you what to do. You can put in extra comments and phrases to make the conversation 'flow'.

(Decide which type of work you would be most interested in. Your address is shown.)

A, You are the manager at the holiday camp. You answer the phone and talk to pupil **B**.
B, Your part is written down. You have to read it. You start by answering the phone.

Appendix F

Instructions for teachers on carrying out the testing

Test format and carrying out
(extract from Teachers' Handbook)

The speaking test comes in three versions. As pupils take the test in pairs (A and B), the material is printed in separate booklets for pupils A and B.

Version one is intended as a practice test, and should be used by all pupils in the class. This is also the version used on the video recording, and pupils can watch this recording following their booklets, preferably after trying out the tasks themselves. As pupils will probably be using the practice material as classroom activity, teachers are recommended to buy enough sets of version one to 'cover' the whole group. (Remember that a package of five pairs of booklets will cover 10 pupils – five 'A's and five 'B's!)

Versions two and three are intended for the 'real test'. Pupils should do either version two or version three. (Teachers should alternate between using these versions.) Versions two and three are not intended to be distributed to the class, but should only be shown to the pupil during the actual test. The same booklets can be reused several times. For this reason, teachers will not normally need more than one pack of five pairs of booklets for versions two and three.

The test is done with pupils in pairs, using test booklets A and B. Three tasks are given, each of the pair having to carry out his or her side of the task, as the booklet indicates. Altogether the test takes about 30 minutes.

The test is intended to be done ideally with a teacher supervising and directing the tasks. Each 'real' test version (two and three) has a set of comprehensive instructions (presented in the Appendix), which should be read out by the supervisor to guide the pupils through the test. It is important to stick fairly closely to this 'script', to ensure that each pupil fulfils his/her role, and that pupils are given the same 'input' from the supervisor. The pupils are also given support from clear instructions in the test booklet, and the use of pictures is widespread. The tasks should be done in the order presented.

Either the teacher can record the whole test on audio-tape or video and evaluate it later, or a second teacher can be involved and evaluation can take place on the spot (either way, it can be an advantage to make a recording, both for evaluation purposes and in order to keep a record).

If this method should prove impractical, particularly if the teacher wishes to test the whole class, the material can be used differently. For example, pupils can be given the material and carry out the test themselves, in a side room, while recording the performance. A third pupil with good language skills should be used to direct the test.

The advantages of using a third person (supervisor) are threefold. Firstly s/he can act as a back-up, keeping the test 'moving', if one or other of the pupils has difficulty in completing his/her side of the task. Secondly, it ensures that pupils are referred to by name as they begin their part of the tasks, which is essential when material is being assessed from a recording. And finally, the third person can have responsibility for the practical side of the recording itself.

The teacher may choose to use a combination of these methods, depending on the maturity of the pupils involved, the class size, and other practical considerations.

Appendix G

Package of forms and checklists for assessment of performance during routine classroom activity

1 Checklist of what to listen for in assessing speaking and listening

the language used

Were the language structures and vocabulary

- appropriate and fully understandable, without many errors?
- on the whole understandable, despite errors and gaps?
- so full of errors and gaps that the message was not easily understood?

Were the language choices

- adventurous, independent and idiomatic?
- independent and idiomatic at times?
- very dependent on the 'given' material/input?

What about

- the pronunciation – was it good, or even understandable?
- the intonation – did it support the message?
- the ability to read aloud?
- the ability to adapt the language according to the social and physical context?
- the ability to speak on the telephone?

When difficulties in communication arose, did the pupil attempt to overcome these in English?

the task done

- Could the pupil carry out the specific tasks required?
- Was the pupil able to take the initiative and 'keep going' without a lot of prompts?
- Did the tone indicate friendliness, politeness and interest?

What about

- the ability to take part in short exchanges?
- the ability to hold the floor for a minute or two?

the listening

- Did the pupil need a lot of simplifying and help in order to understand what was said?
- Was the pupil good at identifying his/her own difficulty in understanding, and actively seeking help to put this right?

2 Teacher's assessment form – speaking and listening

Teachers are referred to the checklist for speaking and listening for a fuller description of what to look for when filling in this form.

Award points out of 5 as follows:
5 = very good
4 = good
3 = adequate
2 = rather weak
1 = very weak

Pupil .....

task .....

date .....

contribution and ability to keep going	[ ]
intonation and manner	[ ]
pronunciation	[ ]
language choices	[ ]
language accuracy	[ ]
structuring	[ ]
listening (understanding)	[ ]
listening (clearing up misunderstanding if necessary)	[ ]

3 Pupil's self-assessment – speaking and listening

name ..... class .....

Think about the activity you've just done, and answer these questions as well as you can. Perhaps your group or partner can help you.
Put a cross in one of the boxes. Three of them are named – *no, usually, always.* The other two are 'in between' these – *rarely, most of the time.*

1 When I was talking was I able to **'keep going'**? (without help, or stopping a lot)

no [ ] [ ] usually [ ] [ ] always [ ]

2 Was my **pronunciation** good enough for the others to understand me?

no [ ] [ ] usually [ ] [ ] always [ ]

3 Did I know enough **words and grammar** to say what I wanted?

no [ ] [ ] usually [ ] [ ] always [ ]

4 When I didn't know a word, was I able to **explain what I meant** in English?

no [ ] [ ] usually [ ] [ ] always [ ]

5 Did I remember to be **friendly and polite, and show interest in the others**?

no [ ] [ ] usually [ ] [ ] always [ ]

6 Did I **understand** what the others said?

no [ ] [ ] usually [ ] [ ] always [ ]

7 If I didn't understand at first, was I able to **ask for help**, in English, so that I understood in the end?

no [ ] [ ] usually [ ] [ ] always [ ]

Appendix H

Self-assessment

Self-assessment: Speaking

How well would you manage these tasks in English?

1 Discussing with a friend which equipment you should take on holiday

I would manage it
1 ❑ *very badly* 2 ❑ *rather badly* 3 ❑ *quite well* 4 ❑ *well* 5 ❑ *very well*

2 Talking for a couple of minutes about your neighbourhood

I would manage it
1 ❑ *very badly* 2 ❑ *rather badly* 3 ❑ *quite well* 4 ❑ *well* 5 ❑ *very well*

3 Answering a few questions about yourself (name, favourite food, etc.)

I would manage it
1 ❑ *very badly* 2 ❑ *rather badly* 3 ❑ *quite well* 4 ❑ *well* 5 ❑ *very well*

4 Phoning a station to ask for train times

I would manage it
1 ❑ *very badly* 2 ❑ *rather badly* 3 ❑ *quite well* 4 ❑ *well* 5 ❑ *very well*

Appendix I

Smallwords analysed groupwise for overall use and turn position

smallword	group	occurrences	mean pr pupil	user proportion	turn initial	turn-internal	turn-final	loner
right	NS	101	5.8	17/18	55	14	1	31
	NoA	2	0.1		2/19	0 1	1	0
	NoB	0	0	0 /24	0	0	0	0
all right	NS	22	1.2	12/18	18	1	0	3
	NoA	8 0.	4	1/19	2	2	0	4
	NoB	8	0.3	4/24	4	0	0	4
okay	NS	52	2.9	16/18	22	11	1	18
	NoA	132	14.5	17/19	52	14	2	62
	NoB	77	3.2	18/24	51	9	2	15
well	NS	74	4.1	18 /18	48	24	0	2
	NoA	47	2.3	13/19	32	15	0	0
	NoB	16	0.7	9/24	12	4	0	0
oh	NS	15	0.8	14/18	6	8	0	0
	NoA	29	1.4	18/19	17	6	1	5
	NoB	25	1.0	15/24	15	4	0	6
ah	NS	20	1.1	10/18	15	5	0	0
	NoA	6 0.	3	6/19	4	1	0	
	NoB	2	0.1	2/24	1	1	0	0
I think	NS	29	1.6	13/18	7	22	0	0
	NoA	71	3.7	18/19	16	46	7	2
	NoB	56	2.3	18/24	24	26	6	0
you see	NS	4	0.2	3/18	0	4	0	0
	NoA	1	0.05	1/19	1	1	0	0
	NoB	0	0	0/24	0	0	0	0
you know	NS	4	.2	3/18	0	4	0	0
	NoA	2	.1	2/19	0	0	2	0
	NoB	3	.1	2/24	0	2	1	0
I see	NS	1	.05	1/18	0	1	0	0
	NoA	2	0.1	1 /19	1	1	0	0
	NoB	2	.1	1/24	0	0	0	2
I know	NS	2	.05	2/18	2	2	0	0
	NoA	3	.1	3/19	0	0	1	2
	NoB	0	0	0/24	0	0	0	0
I mean	NS	2	1.1	2/18	0	2	0	0
	NoA	0	0	0/19	0	0	0	0
	NoB	1	.04	1/24	0	1	0	0
just	NS	102	5.1	17/18	5	96	1	0
	NoA	42	2	17/19	7	35	0	0
	NoB	31	1.2	15/24	2	27	2	0
sort/kind of	NS	24	1.3	7/18	1	23	0	0
	NoA	5	.25	5/19	1	4	0	0
	NoB	0	0	0/24	0	0	0	0
like	NS	46	2.7	15/18	3	43	0	0
	NoA	3	.1	2/19	0	3	0	0
	NoB	1	0	2/24	0	1	0	0
a bit	NS	17	.8	8/18	1	16	0	0
	NoA	9	.45	3/19	0	7	2	0
	NoB	6	.25	5/24	1	5	0	0
or somethg	NS	17	1	7/18	0	12	5	0
	NoA	24	1.2	8/19	0	11	13	0
	NoB	14	.6	8/24	0	5	9	0

Appendix I

smallword	group	occurrences	mean pr pupil	user proportion	turn initial	turn-internal	turn-final	loner
not really	NS	4	.6	8/18	2	2	0	0
	NoA	5	.3	5/19	0	5	0	0
	NoB	2	.1	1/24	0	1	0	1
and	NS	14	.6	8/18	0	12	2	0
everythg	NoA	2	.1	2/19	0	2	0	0
things/	NoB	2	.1	2/24	0	2	0	0
stuff/ that								
total	NS	550	30,5555556	9.5/18 (53%)	185	302	10	54
	NoA	393	20,6842105	6.5/19 (32%)	133	155	29	76
	NoB	246	10,25	5.7/24 (24%)	110	84	20	28
range	NS	19			11	16	5	4
	NoA	17			7	11	8	4
	NoB	15			6	10	5	5
pr 10000	NS	445,38019	24,743344		149,8097	244,554	8,098	43,73
	NoA	279,39713	14,705112		94,55424	110,195	20,62	54,03
	NoB	235,02436	9,79268176		105,0922	80,2522	19,11	26,75
smw wd counts								
	NS	684						
	NoA	524						
	NoB	341						
filled pauses								
	NS	348						
	NoA	845						
	NoB	815						
‘proper’ words								
	NS	11317						
	NoA	12697						
	NoB	9311						

Appendix J

Smallwords as (micro)signals

code	group	tokens	1a take turn	1a hold turn	1a yield turn	1b oblique	2a mode change	2b mid break	3 change of state	4a vague, etc.	5a acknowledge	5b appeal
right	NS	101	55	14	1	0	17	0	0	0	39	1
	NoA	2	0	1	1	0	0	0	0	0	1	1
	NoB	0	0	0	0	0	0	0	0	0	0	0
all right	NS	22	18	1	0	0	1	0	0	0	4	1
	NoA	8	2	2	0	0	1	0	0	0	4	0
	NoB	8	4	0	0	0	0	0	0	0	5	0
okay	NS	52	22	11	1	0	1	1	0	0	21	2
	NoA	132	52	14	2	0	16	0	0	0	72	2
	NoB	77	51	9	2	0	15	2	0	0	25	1
well	NS	74	48	24	0	10	26	5	0	6	3	0
	NoA	47	32	15	0	13	7	7	0	2	0	0
	NoB	16	12	4	0	2	5	2	0	0	0	0
oh	NS	15	6	8	0	0	0	6	5	1	0	0
	NoA	29	17	6	1	0	1	2	10	3	5	1
	NoB	25	15	4	0	0	0	1	5	0	10	1
ah	NS	20	15	5	0	2	0	1	3	0	4	0
	NoA	6	4	1	0	0	1	1	0	0	1	0
	NoB	2	1	1	0	0	0	0	0	0	1	0
I think	NS	29	7	22	0	0	0	0	0	29	0	0
	NoA	71	16	46	7	0	0	0	0	71	0	0
	NoB	56	24	26	6	0	0	0	0	56	0	0
I mean	NS	2	0	2	0	0	0	2	0	0	0	0
	NoA	0	0	0	0	0	0	0	0	0	0	0
	NoB	1	0	1	0	0	0	1	0	0	0	1
you see	NS	4	0	4	0	0	0	0	0	0	0	4
	NoA	1	1	1	0	0	0	0	0	0	0	1
	NoB	0	0	0	0	0	0	0	0	0	0	0
you know	NS	4	0	4	0	0	0	0	0	0	0	4
	NoA	2	0	0	2	0	0	0	0	0	0	2
	NoB	3	0	2	1	0	0	0	0	0	0	3

code	group	tokens	1a take turn	1a hold turn	1a yield turn	1b oblique	2a mode change	2b mid break	3 change of state	4a vague, etc.	5a acknowledge	5b appeal
I see	NS	1	0	1	0	0	0	0	1	0	0	0
	NoA	2	1	1	0	0	0	0	2	0	2	0
	NoB	2	0	0	0	0	0	0	2	0	1	0
I know	NS	2	2	2	0	0	0	0	0	7	1	0
	NoA	3	0	0	1	0	0	1	0	0	2	0
	NoB	0	0	0	0	0	0	0	0	0	0	0
just	NS	102	5	96	1	0	0	0	0	102	0	0
	NoA	42	7	35	0	0	0	0	0	42	0	0
	NoB	31	2	27	2	0	0	0	0	31	0	0
sort/kind of	NS	24	1	23	0	0	0	0	0	24	0	0
	NoA	5	1	4	0	0	0	0	0	5	0	0
	NoB	0	0	0	0	0	0	0	0	0	0	0
like	NS	46	3	43	0	0	0	0	0	46	0	0
	NoA	3	0	3	0	0	0	0	0	3	0	0
	NoB	1	0	1	0	0	0	0	0	1	0	0
a bit	NS	17	1	16	0	0	0	0	0	17	0	0
	NoA	9	0	7	2	0	0	0	0	9	0	0
	NoB	6	1	5	0	0	0	0	0	6	0	0
or	NS	17	0	12	5	0	0	0	0	17	0	0
something	NoA	24	0	11	13	0	0	0	0	24	0	0
	NoB	14	0	5	9	0	0	0	0	14	0	0
not really	NS	4	2	2	0	0	0	0	0	4	0	0
	NoA	5	0	5	0	0	0	0	0	5	0	0
	NoB	2	1	1	1	0	0	0	0	2	0	0
and	NS	14	0	12	2	0	0	0	0	14	0	0
things	NoA	2	0	2	0	0	0	0	0	2	0	0
etc.	NoB	2	0	2	0	0	0	0	0	2	0	0
								0				
total	NS	550	185	302	10	12	45	15	9	267	72	12
	NoA	393	133	154	29	13	26	11	12	166	87	7
	NoB	345	111	88	21	2	20	6	7	112	42	6
								0				
range	NS	19	12	15	5	2	1	5	2	9	4	5
	NoA	17	7	10	8	1	1	5	1	8	4	5
	NoB	15	6	10	5	1	1	3	1	6	4	1
pr 10000	NS	445,38	149,80	244,55	8,10	9,72	36,44	12,15	7,29	216,21	58,30	9,72
	NoA	279,40	94,55	109,48	20,61	9,24	18,48	7,82	8,53	118,02	61,85	4,98
	NoB	329,61	106,05	84,07	20,06	1,91	19,11	5,73	6,69	107,00	40,13	5,73

Index

C

D

E

F

G

H

I

S

T