Testing the Spoken English of Young Norwegians

A study of test validity and the role of ‘smallwords’ in contributing to pupils’ fluency
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
I Oral language assessment and conversation analysis

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Unpublished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
## Acknowledgements

## Series Editor’s notes

## Chapter 1

**Introduction**

- Test validation 1
- Fluency and smallword use 2
- The test 3
- Research questions 3
- Data and methods 4
- Organisation of the book 5

## PART ONE: TEST VALIDATION

### Chapter 2

**Test validation**

- Validation – an overview 10
- Content validation 12
- Face validation 13
- Response validation 14
- Washback validation 17
- Consequential validation 18
- Criterion-related validation 18
- Reliability 20
- Test bias 22
- Construct validation 23
- The narrower view of construct validity 24
- The broader, unifying, view of construct validity 25
- Threats to validity summarised 27
- A unified framework for validation 28
- Six central aspects of validity 29
- A validation framework 30
- Towards the validation process 32
Chapter 3
Communicative language ability 33
Towards a model of communicative competence 34
Models of communicative competence reviewed 35
A suitable model of CLA 39
Describing the domain of CLA 43
Speaking 43
The situation of the testees 46
Operationalising components of CLA 49
Operationalising microlinguistic ability 50
Operationalising textual ability 51
Operationalising pragmatic ability 52
Operationalising strategic ability 53
Some conclusions on CLA and the significance of smallwords 54
Summary 55

Chapter 4
Validation of the test ‘as it stands’ 58
The aims and purposes of the EVA testing 59
Speaking test specifications 59
Specifications for elicitation procedures 60
Specifications for scoring procedures 62
The validation process 65
The CONTENT aspect of validity 66
The SUBSTANTIVE aspect of validity 71
The STRUCTURAL aspect of validity 74
The GENERALISABILITY aspect of validity 83
The EXTERNAL aspect of validity 84
The CONSEQUENTIAL aspect of validity 86
Summary and conclusions 88
Conclusion on the extent to which the model of CLA is represented in the test 88
Conclusions on the validity of the test 90

Chapter 5
Validation based on scoring data 96
Data and methods 96
The a posteriori validation process 99
The EXTERNAL aspect 100
The CONTENT aspect: test bias with respect to gender 103
Generalisability 104
Inter-rater reliability 104
PART TWO: FLUENCY AND SMALLWORD USE

Chapter 6
Fluency and smallwords – making the connection
Fluency
Pinning down fluency
Identifying elements of fluency
A language of fluency?
Fluency summarised
Forging a link between smallwords and fluency
Smallwords in other people’s books
Smallwords and fluency in relevance theory terms
The essence of relevance theory
Proposing a role for smallwords in relevance theory
The work of smallwords in optimalising fluency
Levelt’s perspective: speech production and fluency
A framework for analysing smallword signals
Summary

Chapter 7
Smallwords and other fluency markers: quantitative analysis
The approach
The data
The smallwords
Hypotheses and research questions
Method
Findings on temporal variables
Filled pausing
Mean length of turn
Conclusions on temporal variables
Findings on smallwords
General smallword use: quantity and distribution
Range and variety in smallword use
Smallword use summed up
Smallwords and filled pausing
Summary
## Chapter 8
### The signalling power of smallwords

- The approach 184
  - Data, hypotheses and research questions 185
  - Method 185
- Defining and analysing evidence that smallwords are used to send signals 188
- Expressing the communicative intention 188
  - Signalling whether the speaker intends to take, hold or yield the turn 189
  - Signalling an oblique response 192
- Pointing to the context for interpretation 194
  - Signalling a break with the initial context created by the previous speaker (‘mode changing’) 194
  - Signalling a mid-utterance break with context created by the speaker’s own immediately preceding speech 196
- Indicating the cognitive effect of the previous utterance 200
  - Signalling a cognitive change of state, resulting from the previous utterance 201
  - Indicating the degree of vagueness or commitment: Signalling a softening of the impact of the message, or ‘hedging’ 204
  - Learner-favoured hedges 208
  - Learner-underused hedges 209
  - Conclusions on hedging 213
- Indicating the state of success of communication 213
  - Signalling the acknowledgement of smooth communication 213
  - Signalling an appeal to the listener to confirm or assist smooth communication 216
- Summary 218

## Chapter 9
### The smallword user

- Variation in smallword use 224
  - Gender 224
  - Task 226
- The acquisition of smallwords 229
- The implications of the findings for language education 232
  - Implications for assessment 233
  - Implications for teaching and learning 237
- Summary 239
## CONCLUSION

**Chapter 10**

**Conclusion**

<table>
<thead>
<tr>
<th>Sections</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>The research questions</td>
<td>243</td>
</tr>
<tr>
<td>The findings</td>
<td>244</td>
</tr>
<tr>
<td>Theoretical findings</td>
<td>244</td>
</tr>
<tr>
<td>Empirical findings</td>
<td>248</td>
</tr>
<tr>
<td>A small word in conclusion</td>
<td>254</td>
</tr>
</tbody>
</table>

**Glossary**

| Pages | 255 |

**References**

| Pages | 259 |

**Appendices**

| Pages | 267 |

**Index**

| Pages | 295 |
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

2 Test validation
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