



Developing a validation process for an adaptive computer-based spoken English language test.

UNDERHILL, Nic.

Available from the Sheffield Hallam University Research Archive (SHURA) at:

<http://shura.shu.ac.uk/20468/>

A Sheffield Hallam University thesis

This thesis is protected by copyright which belongs to the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Please visit <http://shura.shu.ac.uk/20468/> and <http://shura.shu.ac.uk/information.html> for further details about copyright and re-use permissions.

Sheffield Hallam University
Learning and IT Services
Adsett Centre City Campus
Sheffield S1 1WB

101 653 022 6



REFERENCE



ProQuest Number: 10701115

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



ProQuest 10701115

Published by ProQuest LLC (2017). Copyright of the Dissertation is held by the Author.

All rights reserved.

This work is protected against unauthorized copying under Title 17, United States Code
Microform Edition © ProQuest LLC.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Developing a validation process for an adaptive computer-based spoken English language test

Nic Underhill

A thesis submitted in partial fulfilment of the
requirements of Sheffield Hallam University for the
degree of Doctor of Philosophy

July 2000

Abstract

This thesis explores the implications for language test validation of developments in language teaching and testing methodology, test validity and computer-based delivery. It identifies a range of features that tests may now exhibit in novel combinations, and concludes that these combinations of factors favour a continuing process of validation for such tests. It proposes such a model designed around a series of cycles drawing on diverse sources of data.

The research uses the Five Star test, a private commercial test designed for use in a specific cultural context, as an exemplar of a larger class of tests exhibiting some or all of these features. A range of validation activities on the Five Star test is reported and analysed from two quite different sources, an independent expert panel that scrutinised the test task by task and an analysis of 460 test results using item-response theory (IRT). The validation activities are critically evaluated for the purpose of the model, which is then applied to the Five Star test.

A historical overview of language teaching and testing methodology reveals the communicative approach to be the dominant paradigm, but suggests that there is no clear consensus about the key features of this approach or how they combine. It has been applied incompletely to language testing, and important aspects of the approach are identified which remain problematic, especially for the assessment of spoken language. They include the constructs of authenticity, interaction and topicality whose status in the literature is reviewed and determinability in test events discussed.

The evolution of validity in the broader field of educational and psychological testing informs the development of validation in language testing and a transition is identified away from validity as a one-time activity attaching to the test instrument towards validation as a continuing process that informs the interpretation of test results.

In test delivery, this research reports on the validation issues raised by computer-based adaptive testing, particularly with respect to test instruments such as the Five Star test that combine direct face-to-face interaction with computer-based delivery.

In the light of the theoretical issues raised and the application of the model to the Five Star test, some implications of the model for use in other test environments are presented critically and recommendations made for its development.

Contents

Chapter	Page
Chapter 1	Introduction: statement of issues and problems 1
Chapter 2	Literature review 1: The evolution of language teaching and testing 8
Chapter 3	Literature review 2: Validity in language testing; testing spoken language; adaptive testing and item response theory..... 61
Chapter 4	The instrument: the Five Star test and comparisons with other tests 130
Chapter 5	Data collection and research methodology 172
Chapter 6	Analysis and discussion of results 214
Chapter 7	Discussion of critical thinking..... 300
Chapter 8	Derivation of theoretical model for continuous validation 335
Chapter 9	Application of the model for validation of Five Star test..... 385
Chapter 10	Conclusions, reflections and implications 403
Bibliography.....	419
Appendices.....	433

List of Figures

Figure	Page
Figure 1 Model of spoken language ability used in the CASE project	41
Figure 2 Bachman's 1990 components of language competence model	43
Figure 3 Schematic diagram of Bachman's (1990) model of validity	79
Figure 4 Language production scale (Palmer, 1981)	115
Figure 5 QUEST output of item-ability map	284
Figure 6 Individual IRT learner map for candidate 001.....	286
Figure 7 Individual IRT learner map for candidate 356.....	288
Figure 8 Diagrammatic representation of model for continuing test validation	365
Figure 9 Common framework for pilot and main test cycles	369
Figure 10 Short- and long-term stakeholder feedback	376

List of Tables

Table		Page
Table 1	Task-driven and construct-driven performance assessment.....	48
Table 2	Summary of major models of validity	88
Table 3	Distinctive features of Five Star computer platform.....	140
Table 4	Analysis of Five Star tasks and scoring criteria.....	143
Table 5	Summary of key communicative features	152
Table 6	Other current oral test formats	157
Table 7	Panel stage 1 <i>pro formas</i> : language skills definitions.....	183
Table 8	Panel stage 1 <i>pro formas</i> : language skills allocations.....	185
Table 9	Routes through Five Star tasks for panel stage 1	186
Table 10	Panel stage 2 <i>pro formas</i> : language skill and level allocations.....	188
Table 11	External rating scale used for panel judgements of proficiency	191
Table 12	Panel rubric for interaction strategy exercise.....	193
Table 13	Basis for panel consensus on skills allocation	216
Table 14	Non-significant consensus on skills allocation.....	217
Table 15	Panel judgements on skills allocation for each task.....	218
Table 16	Frequency of skill allocations across all tasks	221
Table 17	Frequency of skill combinations across all tasks.....	222
Table 18	Panel percentage judgements of skills underlying each task	223
Table 19	Panel ratings for proficiency levels required for each task	227
Table 20	Sample of panellists' comments and suggestions: first 15 tasks	231
Table 21	Panel stage 3 <i>pro formas</i>	235
Table 22	Panel judgements of candidates' interaction strategies	239
Table 23	Summary of interaction strategies by panel sub-groups	249
Table 24	Percentage of use of interaction strategies, by panel sub-groups.....	250
Table 25	Summary of interaction contribution and proficiency ratings	254
Table 26	Video test proficiency ratings	255
Table 27	Summary IRT statistics for the item (task) estimates	264
Table 28	Summary IRT statistics for the case (candidate) estimates.....	265
Table 29	QUEST output for individual item estimates	267
Table 30	IRT statistics for misfitting tasks.....	272

Table 31	QUEST output for individual case (candidate) estimates	274
Table 32	Comparison of item difficulty estimates between data sets	290
Table 33	Correlations between IRT and panel estimates of task difficulty.....	292
Table 34	Panel consensus for misfitting tasks.....	293
Table 35	Summary of preliminaries to test validation model	338
Table 36	Checklist of test characteristics for the validation model	349
Table 37	Preliminary activities and sources of data for Five Star test.....	387
Table 38	Validation activities and sources of data for Five Star test.....	389

Chapter 1 Introduction: statement of issues and problems

- 1.0 Introduction
- 1.1 Aims
- 1.2 Introduction to the instrument
- 1.3 Major issues

1.0 Introduction

This chapter briefly introduces the aims of this research, the test instrument on which the data were collected, and summarises the issues which it raises which are explored in more detail in later chapters.

The research grew out of a consultancy project to carry out a critical review of a new computer-based test of English language proficiency. A major aim of the consultancy was the preliminary validation of the pilot form of the test, in order to inform its subsequent development, and some of the data collected was first analysed within the scope of that project. However, the methodology adopted was constrained by the nature of the test and the resources and timescale available for the consultancy, and it became clear that the traditional approach to test validation could not provide a fully satisfactory account. Specifically, the Five Star test is both adaptive (different subjects will take different routes through the test, and be exposed to different tasks), and scored on a partial credit basis (each task is scored on a three-point scale, rather than dichotomously, as right or wrong). At the same time, it is a direct test of spoken

language that exhibits many features of the communicative approach to language testing.

The Five Star test is introduced in section 1.2 below and described in more detail in chapter four.

1.1 Aims

These are the aims of the research.

Aim 1 is to design a model appropriate for an adaptive test of spoken language proficiency that allows for the validation to become a recurrent process as the test evolves rather than a single procedure. The development of the model draws on a number of strands: the evolution of language testing methodology, new approaches to test validation, contrasting sources of data and the evolution of a new adaptive computer-based language tests. The research explores the use of appropriate research methods to analyse the expert panel and test record data sets in a way that supports a continuing cycle of data collection, analysis and interpretation.

Aim 2 is to identify the distinctive features of the communicative approach to language teaching and testing and to discuss their implications for the model for continuing validation. The nature of interaction in particular is viewed as a criterial feature of direct communication.

Aim 3 is to try out the model to validate the 'Five Star' computer-based test of language proficiency within its immediate social and cultural context. This builds on the initial 'expert panel' data set collected during the consultancy project, and subjects it to further analysis; and adds a second large dataset, using the item response theory (IRT) approach to the analysis of test data, to complement and triangulate the expert panel data. The Five Star test developer wrote at an early stage of the prototype "... the open architecture aspect of computerisation would also mean that the existing aspects of the [Five Star] test could be easily modified and new ones included" (Pollard, 1994:43) and the validation process needs to be able to cope with this.

Aim 4 is to discuss the implications of aims 1, 2 and 3 to explore a procedure that can be applied elsewhere for the validation of language proficiency tests that share some or all of these key features. This seeks to identify the components of the continuing validation process which can be transferred to other testing contexts, specifically, where the continuing validation of 'direct' oral tests needs to be subjected to greater scrutiny. The use of the model for the validation of a continuously changing test is envisaged.

Aim 5. As a result of aims 1, 2, 3 and 4, aim 5 is to contribute to and enrich, at a theoretical level, the academic debate on issues surrounding language test validation. Review of recent literature suggests that while the validation of adaptive tests has become a current issue of debate in language testing, with use of the item-response theory (IRT) in particular, such tests are almost always confined to objectively marked, single-correct-answer only items.

The applicability of such validation procedures for non-dichotomous items in a test procedure such as Five Star, based on live human interaction with an interlocutor (see Appendix 1 Glossary), represents an extension of this work. This integration of human-mediated interaction and assessment with computer-based adaptivity based on a theoretical foundation of communicative methodology is unique in a language proficiency test, and it is the elaboration of procedures for the continuing validation of this novel combination of features that represents the contribution to theoretical debate on language test validation.

1.2 Introduction to the instrument

The Five Star test is a computer-based test of English language proficiency; its name derives from the scale of one (star) to five (stars) on which scores are reported. It was developed by Saudi Development and Training (SDT), a subsidiary of British Aerospace, specifically for use in the Saudi Arabian context to screen young adult applicants for work positions and training programmes, and was first reported in Pollard (1994) and Pollard and Underhill (1996). The piloting of the prototype test started in 1993. It has now been used with over 1000 members of the target population - male, young adult Saudis seeking to enter the job market - mainly within the parent organisation, but also on a limited basis with other companies.

The test offers a unique profile in being a) adaptive, b) computer-based but interviewer-administered, and c) heavily dependent on interaction with a live interlocutor. It is administered to a single participant at a time by a single interviewer, and consists of a

selection of the tasks presented by the computer but mediated and scored by the interviewer. The act of scoring each task via the keyboard invokes the computer algorithm which uses the information to select the next task for the test. Typically, a test administration will involve between eight and 15 tasks and last anywhere between 10 and 40 minutes. Tests with candidates of higher levels of language proficiency involve more tasks and take longer than those for more elementary users of English.

In the very wide range of English language tests in use around the world, the Five Star test can be characterised as follows.

It is a *proficiency* test. It aspires to test a candidate's overall proficiency in the English language, and is not based on a specific syllabus or achievement or progress in a specific course of study. It can in principle be equally applied to a candidate with a limited command of English and to a candidate with a high level of proficiency.

It is a test of *English for Speakers of Other Languages*; crudely speaking, it is aimed at non-native rather than native speakers of English. The pilot version is designed exclusively for candidates whose mother tongue is Arabic.

It is *live* and *direct*; there is immediate 'real time' verbal interaction with an interlocutor who also assesses performance on each task (the 'interviewer'). The dialogue is not scripted or predictable, except in general terms.

It is *adaptive*, in two respects. Firstly, the algorithm underlying the computer program varies the selection of tasks to present to each candidate according to their performance

on previous tasks. Secondly, both the interviewer and the candidate may, in the nature of live interaction, take the initiative to alter the focus or even the topic of the conversation.

It is *task-based*. The test items consist of complex open-ended (for the most part) activities that require the demonstration of a language skill or combination of skills, rather than single questions to which there is a single correct answer. Performance on each task is scored on a three-point scale.

It is *culture specific*. The test was designed specifically for use within the Kingdom of Saudi Arabia, and makes extensive use of the Arabic language, in both script and digitised sound.

It is *topical*. Some of the tasks require reference to recent local and regional knowledge of the world in the social and geographical context of that culture.

The Five Star test, and these characteristics, are described in more detail in chapter four. On a terminological note, the person taking the test is referred to as the 'candidate', and the person administering the test as the 'interviewer'. However, the interviewer combines the roles of 'interlocutor' and 'assessor' and these terms are used when the respective roles are being examined. These terms are defined in the glossary.

1.3 Major issues

The major issues raised are these.

- a) How can recent thinking on test validation be applied to an adaptive test? How can the IRT model be used? What are the advantages and disadvantages of an expert panel and IRT as sources of validation evidence?
- b) How can a validation process allow for the implications of the communicative methodology for the continuing development of an adaptive test?
- c) How can the IRT model be applied in a test where the questions are for the most part open-ended tasks evaluated by a human interlocutor on a three-point partial-credit scale, rather than the more conventional right-or-wrong objective items?
- d) How do we deal with the concept of interaction that is central to a live oral test? Is it a construct that can be operationalised and measured as a variable for each individual? If it is not an individual trait, is it a feature that attaches to the event rather than to the person? How can its impact on the test scores be described?

- 2.0 Introduction
- 2.1 Language teaching and testing
- 2.2 The classical approach
- 2.3 The Direct method
- 2.4 The scientific revolution
- 2.5 The performance model
- 2.6 The oral and situational approaches
- 2.7 Learning and acquisition
- 2.8 Humanistic perspectives
- 2.9 The communicative approach: introduction
- 2.10 The communicative approach: British and American approaches
 to communicative testing
- 2.11 Problems with componential models of communicative
 competence
- 2.12 Work sample and cognitive approaches to assessment
- 2.13 The communicative approach: individualisation, task-based
 learning and authenticity
- 2.14 Discussion and summary

2.0 Introduction

This chapter gives an overview of the development of English language teaching methodology, and in particular picks out the parallels between teaching and testing

methodology. Key characteristics of the Five Star test are traced back to their origins in the historical development of language testing and their origins identified in teaching methodologies, in particular the communicative approach. The following chapter then looks specifically at test validation issues in more detail.

2.1 Language teaching and testing

Current approaches to the validation of language tests, and their application to adaptive and computer-based test technologies, are the culmination of a long history of development, and this chapter summarises that evolutionary process to identify the pedagogical and methodological sources of the present-day test characteristics.

A summary review of developments in linguistics, learning theory, language teaching and testing over the last fifty years shows how the relationship has grown more complex with the development of theoretical models in each area. For the purpose of this review, we can identify a series of apparently discrete developments in teaching methodology, but in reality certain themes recur and recombine in new contexts (a point made in both Howatt (1984) and Richards and Rodgers (1986), two of the principal sources on which the following summaries draw).

While the evolution of language teaching and testing theory has evolved in parallel in the United States on the one hand and Britain and Europe on the other, with each developments on each side informing the other, there have been nonetheless distinct differences in approaches. Spolsky (1975), summarised in Valette (1977), identified

three phases of language testing: pre-scientific, psychometric-structuralist and psycholinguistic-sociolinguistic. The dominant communicative approach can be considered either as falling within the third of these phases or as a distinct fourth phase (Weir, 1990). A rough congruence can usefully be drawn between these phases of testing and related developments in teaching without a complete match being sought.

2.2 The classical approach

Until the late 19th century, the classical languages of Latin and Greek were held up as ideal languages in the West, and the terminology used to describe their syntactical and morphological forms, such as case and tense, was also used to analyse modern languages. Any difficulty in the application of this terminological framework was an indication of the modern language's degeneracy, compared to the classical ideal. As well as a desirable accomplishment in its own right, the practice of learning and analysing a language was believed to be a valuable intellectual exercise.

The associated language teaching methodology was grammar-translation. It emphasised accuracy, in the written language especially; knowing a language involved mastery of a set of rules. There was no direct link with the real world of experience, as the goal of instruction was to master a closed linguistic system. This was achieved by memorising whole chunks of the system, from which the internal contrasts could then be deduced. Areas of specific competence were typically grammar, vocabulary and stylistic rules for writing and to a limited extent speaking: the rule against splitting infinitives, for

example, being a late 19th century invention. There being little purpose in learning to speak a dead language, speaking skills were subordinated to reading skills.

Learning techniques commonly used appropriate to a 'language as knowledge' paradigm were the deductive application of sets of rules combined with rote learning of vocabulary and the numerous exceptions to those rules. Although every other methodology since has been in some way a reaction against grammar translation, the teaching and testing of language as a system of knowledge still remains a significant influence today, particularly in state sector education systems:

It is still used in situations where understanding literary texts is the primary focus of foreign language study and there is little need for a speaking knowledge of the language. Contemporary texts for the teaching of foreign languages at college level often reflect Grammar Translation principles. These texts are frequently the products of people trained in literature rather than in language teaching or applied linguistics. Consequently, though it may be true to say that the Grammar Translation Method is still widely practised, it has no advocates. It is a method for which there is no theory. There is no literature that offers a rationale or justification for it or that attempts to relate it to issues in linguistics, psychology or educational theory. (Richards and Rogers, 1986:4)

The classical approach: testing

Testing followed the explicit characteristics of the teaching method, with a focus on translation; reading and writing; deductive mastery of the rules; accuracy rather than fluency; and an appreciation of its stylistic features rather than its use for communication. Even where more communicative teaching techniques have entered the classroom today, with a substantial timelag between innovations in teaching and testing and the resource disincentive to introduce more labour-intensive testing communicative methods, it is common to find language tests in use with exactly these grammar translation characteristics. This is typically true of ministries of education where the inertia of long-established pedagogic traditions militates against change, but even in

such cases, there is growing recognition that the tests are failing the students on at least two counts. Firstly, the tests do not accurately reflect candidates' performance in the language skills that employers and higher education institutions now require, and secondly, they exert a strong washback effect on classroom teaching. As a result, the skills are neither taught nor tested.

2.3 The Direct method

The direct method was not so much a specific method as a general approach, grouping together a series of methods, sometimes called 'natural methods', that had in common a sharp contrast with the classical approach:

There is little doubt that it was associated in the public mind with Berlitz, and ... Berlitz teachers like Wilfrid Owen used it to describe their work. Nevertheless, Berlitz himself did not, but preferred to stick to his own 'brand name'... The most reasonable explanation of [the origin of the name 'direct method'] is .. that nobody invented the term, but that it 'emerged'...as a useful generic label to refer to all methods of language teaching which adopted the monolingual principle as a cornerstone of their beliefs (Howatt, 1984:207-8)

The origin of the term 'natural method' is more straightforward: "It overemphasized and distorted the similarities between naturalistic first language learning and classroom foreign language learning ..." (Richards and Rogers, 1986:10)

Another element these different methods shared was an emphasis on the primacy of the spoken language. "*Interaction* is at the heart of natural language acquisition, or *conversation* as Lambert Sauvœur called it when he initiated the revival of interest that led eventually to the Direct method" (Howatt, 1984:192). Interaction remains both central and problematic to current methodology.

While it was clear what the various methods grouped under the 'direct' or 'natural' labels were reacting against, it was not so easy to identify what they stood for:

The Direct Method represented the product of enlightened amateurism. It was perceived to have several drawbacks. First, it required teachers who were native speakers or who had native-like fluency in the foreign language. It was largely dependent on the teacher's skill, rather than on a textbook, and not all teachers were proficient enough in the foreign language to adhere to the principles of the method. Critics pointed out that strict adherence to Direct Method principles was often counterproductive, since teachers were required to go to great lengths to avoid using the [students'] native tongue, when sometimes a simple brief explanation in the student's native tongue would have been a more efficient route to comprehension (Richards and Rogers, 1986:10)

The Direct Method was developed further in Britain by Daniel Jones and Harold Palmer at London University in the early years of the 20th century and its central ideas picked up again in the 1960s and 1970s. Palmer emphasized the importance of meaning, for grammatical structures as well as vocabulary, to be presented and associated with elements of the immediate environment, such as objects found in or brought into the classroom.

In the United States, the shortage of sufficiently trained teachers and the "perceived irrelevance of conversation skills in a foreign language for the average American college student" resulted in a study recommending that "a more reasonable goal ...would be a reading knowledge of a foreign language..." and as a result "the emphasis on reading continued to characterize foreign language teaching in the United States until World War II" (Richards and Rogers, 1986:11)

The Direct method: testing

The same disadvantages that militated against the take-up of Direct Method in schools also precluded its widespread application in testing. The lack of competent teachers and the much greater costs of direct oral testing made it simply impractical in public secondary and college systems. This remains true today where decisions are taken by central government to introduce the teaching of a foreign language (typically English) in the primary school, without the provision of extra specialist teachers. As a result, English is often taught by teachers who themselves have only a poor spoken command of the language. It was and is much more successful in private language schools, with relatively small classes of more highly motivated learners, but these are precisely the kind of students who have less need for examination and certification, so that there is less incentive to develop formal assessment procedures.

Spolsky's 'prescientific trend' in language testing (Spolsky, 1975) subsumes both the grammar-translation and direct methods. Summarising this trend in the United States, where the oral emphasis of the direct method never really took hold, Valette says:

There is a lack of concern for statistical analysis, for objectivity, and for test reliability. The tests themselves are mainly written exercises: translation, composition or isolated sentences. In this "elitist" approach to testing, it is felt that the person who knows how to teach is obviously in the best position to judge the proficiency of the students. (Valette, 1977:308)

One reason for the predominance of written tests was the overriding concern for test objectivity and reliability, on which criterion tests such as Five Star involving live interlocutors, open-ended tasks and the elicitation of samples of natural-like speech failed to measure up.

2.4 The scientific revolution

From the 1920s and 1930s the structuralist school of linguistics emerged in America, most commonly associated with Bloomfield (1935). As descriptivists they could only describe what they could observe; they considered that for scientific purposes the mind did not exist because it could not be observed. It was in part a reaction against the 'armchair theorizing' of the classical approach, but systematic and rigorous procedures for language analysis were also necessitated by the fieldwork being undertaken by Bloomfield and his contemporaries into North American Indian languages. Unable to theorize from an armchair, the researcher's complete ignorance of language under study became a virtue, associated with scientific objectivity. Native speaker informants could be used only in restricted ways to inform the analysis. The spoken language was again the primary object of concern, and necessarily the only concern in the study of those native American languages which lacked a written form. The development of anthropological linguistics promoted a relativistic view that our native language affects the way we view the world, and that therefore our understanding of a language cannot be divorced from the contexts of its use. This became known as the Sapir-Whorf hypothesis, from two of its early proponents, and it influenced both mainstream language teaching (e.g. Lado, 1957) and within a single language group the work of sociolinguists such as Hymes (1968) and Labov (1966).

During the Second World War the US military found themselves with a sudden and massive need for language teaching, particularly of 'difficult' (i.e. non Indo-European) languages, and set about filling the gap with customary zeal. They realised the passive 'reading method' then customary in American schools was wholly unsuited to their

requirements for the rapid teaching of an active conversational proficiency, and invited the linguistic experts to adapt the structuralist linguistic paradigm to a teaching context. They produced a language teaching methodology known as the structural approach (e.g. Fries, 1945).

Behaviourist psychology shared with structuralist linguistics the scientific paradigm requiring full observability and transparency, and they combined to produce a more dogmatic version of the structuralist approach known as audiolingualism (e.g. Lado 1964). Learning is a conditioning process, and involves acquiring a set of habits. This necessitated mechanical activities such as repetition, pattern practice, and the manipulation of substitution tables, with a lack of concern for the conveyance of meaning as external reference that made it ideally suited for extensive use in the language laboratory.

Lado's approach to content specification for teaching and testing was very much an atomistic or discrete point approach. Starting with a contrastive analysis of the learner's mother tongue and target language, in different linguistic sub-disciplines (e.g. morphology, syntax, lexis, phonology) and at different levels (e.g. phoneme, morpheme, word and sentence level) an exhaustive list of discrete language items could be generated on which to base mechanical language practice teaching and testing. Although attractive as a mechanistic model reflecting the 'building blocks' paradigm of structuralist linguistics, it suffers from two related practical problems. Firstly, it is in practice extremely difficult to design drills or test items that practise one element at one level only. Secondly, it relies on the combination of building blocks being as predictable and as rule-bound as the basic units themselves. In practice, precise rules become more

complex at the clause and sentence level and it is only quite recently - in the 1980s and 1990s - that any serious attention has been paid to analysis at the level of discourse.

However, remnants of audiolingualism linger on, five decades later, in the American Language Course sold as part of US defence contracts and in the continuing widespread use of pattern practice drill exercises in language laboratories around the world. Although pedagogically outdated, these offer students a high level of exposure to a clear native speaker model of spoken language in contexts where local teachers may not be able to provide this.

The scientific revolution: testing

Spolsky (1975) referred to this as the 'psychometric-structuralist' trend in language testing. The operationalisation of language proficiency for testing purposes in this paradigm is straightforward. There is a single linear dimension of language proficiency, which all tests reflect; some are better than others and, being more reliable, objective tests are better than subjective tests. This was contrasted with the discrete-point approach, which emphasized the differences between what tests measured rather than the similarities.

During this period also statisticians such as Spearman and Pearson developed the statistical tools which have formed the cornerstone of psychometric testing ever since: correlation and regression, factor analysis, internal reliability coefficients and item analysis. For Valette, language testing could be taken seriously as last: "The psychometric-structuralist trend saw the entry of the experts into the field of language testing" (1977:308). Much of the conceptual framework in psychological testing was

directly transferred along with the statistical techniques to language testing, although not without questioning, at least in hindsight:

The application of procedures deriving from psychometric theory in the development of language tests has also met with criticism. Morrow (1979), for example, regards the quantification implied by the use of these methods as inappropriate in the assessment of language proficiency, and Cziko (1983) views psychometric methods as being based too heavily in a norm-referenced interpretation of scores... (Baker, 1997:5)

The impact of psychological measurement led to the positing of general language proficiency and language learning aptitude as unidimensional, linear and stable personal attributes. A heavy reliance on statistical procedures has been the dominant influence: "The foundation for much of classical test theory was provided by Charles Spearman's conception of an observed test score as being composed of a true score plus an error component ... and most of the basic formulae which have proved to be particularly useful in this theory appeared in Spearman's work in the early 1900s" (Baker, 1997:5). The American secondary school system in general, and that of other countries influenced by it, remains heavily biased towards objective forms of assessment and statistical decision-taking, with the result that test methods that are evidently objective are to be preferred to those that are not. Language tests have focused on individual items that are atomistic, with a prescription that they should test only one language point at a time (Lado, 1961).

This unidimensional view of language proficiency, and the preference for objective testing and scoring methods that are susceptible to the rigour of statistical analysis, resulted in a minimal role for oral testing that apparently contradicted the term 'audio-lingual'.

During the past few decades oral language testing has had a great deal in common with physical fitness. Everyone thinks that it is a wonderful idea, but few people have taken time to do anything about it. During the prime period of audio-lingual methodology, for example, the teaching of oral production was the principal classroom objective, but the testing of oral proficiency was almost unknown. (Madsen and Jones, 1981:15).

This has also had a backwash effect on teaching methods, with students being reluctant to devote study time to speaking skills when they know their ability will be tested by objective tests of listening and reading only.

Where, for example, the American Language Course sold as part of the package of defence contracts referred to above is still in use, the mismatch between the skills needed on the one hand, and those taught and tested on the other, has become painfully apparent. Describing the situation at the King Faisal Air Academy in Saudi Arabia in 1994, Al-Ghamdi said:

For a number of years, the American Language Course (ALC) together with its associated ... test, the English Comprehension Level (ECL), have been used as the main language teaching and testing programmes. The ALC is based on a structural syllabus with a heavy emphasis on lexical items, and is audiolingual in approach. The ECL reflects the psychometric-structuralist era as exemplified by Spolsky. It is a 100-item test of listening and reading comprehension ... It soon became obvious that cadets needed an additional range of language skills which were not addressed by the ALC. The increased complaints of Aerospace instructors indicated that there are significant deficiencies in the existing English training programme and that changes should be introduced to bring the English training programme closer to the language skill requirements of the subsequent phases of training (Al-Ghamdi, 1994:5-6).

The results of a needs analysis following a data collection phase were simply that there were major deficiencies in the cadets' speaking ability and Al-Ghamdi goes on to describe the particular problem areas in speaking and to describe how they set about the elaboration of an oral test specifically to fill this gap.

2.5 The performance model

In the structuralist model, the 'rules of grammar' were considered explicit and explicitly teachable until Chomsky (1965) first elaborated the contrast between deep and surface structure. For Chomsky, the rules of grammar are psychologically real, but cannot necessarily be formulated. Although a revolution in hindsight, the 'transformational-generative approach' grew directly out of attempts to extend the explanatory power of the structuralist model to account for irregular inflections and intuitively obvious 'transformational' relations between sentence forms such as active/passive. The ultimate failure of these attempts by American descriptive linguists in the 1940s and 50s such as Harris, Hockett and Nida (Joos, 1957) necessitated the elaboration of a cognitive code approach to explain how we could internalise and use deep-structure rules that could not be explicitly taught.

Chomsky's devastating critique of the behaviourist approach in a review of Skinner's 'Verbal Behaviour' (Chomsky, 1959) brought the mentalist implications of his "deep structure" into stark contrast with the structuralist/behaviourist paradigm that would on principle accept observable data only.

Although associated with a cognitive code approach to the psychology of learning, in direct opposition to the behaviourism of the structuralists, the transformational model remained a linguistic concern, with little direct application to language teaching and testing. Indeed, the location of competence at a 'deep level', accessible only via the surface level of performance, made it necessarily inaccessible to formal teaching or testing. One of the mental constructs that Chomsky elaborated was the Language

Acquisition Device to explain how individual children can apparently be predisposed to learn with such speed and facility the rules of any human language to which they are exposed. This device necessarily operates only on languages to which they are exposed before puberty and of which they can become, in a general sense, 'native speakers'. It does not apply to adult learners of foreign languages, and Chomsky himself saw little relevance in transformational generative grammar for language teaching or testing. In a negative sense, however, it had a considerable impact on teaching methodology, in the sense that it undermined the behaviourist and structuralist bases of audiolingualism. It also laid the foundation, via Hymes (1970) and Morrow (1977, 1979) for the evolution of the communicative approach.

2.6 The oral and situational approaches

In the 1950s and 60s in Europe, the growth in demand for English language teaching as a commercial activity re-awakened interest in the direct method while drawing on a continuing development from Palmer in a direct line through Hornby, Eckersley, Lee, Haycraft and many others who were primarily language teachers rather than linguists or academics. For example, Palmer (1922) was re-issued by Oxford University Press in 1964, Palmer (1932) in 1969. This British re-formulation was characteristically non-ideological; there was certainly an emphasis on the learner's habits and behaviour, but with no theoretical commitment to behaviourism. It was claimed that insistence on the use of English only as the target language in the classroom was intended to teach students to think in English, and to discourage them from constantly translating to and from their mother tongue; more pragmatically, a preference for the use of the target

language only suited the new commercial reality of multi-national groups of learners in each classroom, where it was unrealistic to expect the teacher to speak every mother tongue represented.

The emphasis on the primacy of the spoken rather than the written form (e.g. Billows, 1961) that was inherited from the direct method increasingly matched the needs of language learners, and the term Oral Approach became widespread. For theoretical bases, it drew on both the techniques of audiolingualism and the competence/performance distinction. Pattern practice, in the form of substitution and transformation drills, still had a major role to play, but so too did rule-based learning. However, rather than the classical approach of 'give the rule and require its application', the new direct method invited learners to infer the rule for themselves, then apply it, in the belief that this promotes a more profound internalisation.

What really distinguished the Oral Approach from the Direct Method was a more systematic approach to syllabus content and progression:

An oral approach should not be confused with the obsolete Direct Method, which meant only that the learner was bewildered by a flow of ungraded speech, suffering all the difficulties he would have encountered in picking up the language in its normal environment and losing most of the compensating benefits of better contextualisation in those circumstances (Pattison 1964 in Richards and Rodgers, 1986:34)

This systematic identification and introduction of new language, both grammatical structures and vocabulary, in a step-by-step ordered syllabus favoured the use of specially-written coursebooks that presented the selected language items in a discrete and ordered manner, with a consequent shift of the onus for syllabus design from class teacher to course writer. The presentation of new lexical items in particular was felt to be most naturally made in the everyday contexts in which those words occurred (rather

than, for example in semantic sets, or the decontextualised words lists of the grammar translation approach), and the pursuit of this principle became characterised as the Situational Approach. As early as 1950, Hornby wrote about 'The situational approach in language teaching' (1950) and course texts and methodology books based on this approach published in the 1960s and 70s included 'Situational English' (Commonwealth Office of Education, 1965), 'Situational Dialogues' (Ockenden, 1972) and 'Situational Lesson Plans' (Davies et al., 1975). Howatt notes the situational approach being applied in the Nuffield Foreign Languages Teaching Project (1984: 274)

While the linguistic theory underlying the Situational Approach was a structuralism essentially similar to American structuralism, it also drew on a strain of functionalism in British linguistics (Firth and Halliday) and anthropology (Malinowski and Fraser) that saw context of use as an essential component to the study of language: "The emphasis now is on the description of language activity as part of the whole complex of events which, together with the participants and relevant objects, make up actual situations" (Halliday et al., 1964, quoted in Richards and Rogers, 1986: 35). Rather than being seen as a closed and independent system, language use is a purposeful activity related to everyday life in the real world; we use language to do things. In its strong form, derived from Malinowski's 'context of situation', the meaning of an utterance can only be determined with reference to the cultural and physical context in which it occurs (Malinowski, 1923). This preoccupation with the context of language use remained in the British tradition of English language teaching, and emerges in the concern of the communicative approach to make explicit the functions and appropriate contexts of use of the language that is being taught and tested.

The oral and situational approaches: testing

The specific focus on 'the situation' as the unit for description of the target language domain was reflected in testing as well as in teaching primarily through short role play exercises. A physical and/or a social context was described, such as 'Imagine you're talking to a tourist information office and you want to find accommodation for the night, but you're concerned about the price'. Another participant, typically the interviewer, would play the role of interlocutor in the role-play, and the candidate would be instructed to find out certain information from him or her. Such mini-roleplays were in use in the Cambridge main suite exams into the 1980s and are still used in modified form in their CCSE exams (UCLES, 1995). Typically these situations were kept very short, and the rubric to the candidate might simply be 'What would you say if ...' with no time for preparation and little contextualisation to tap into strategic or interactional competence.

The emphasis on oral fluency in lifelike contexts - performing in the language, rather than demonstrating competence of it - anticipated the construct of communicative competence and the discussion about how to assess it. The specific focus on oral work necessitated reviewing the assumption of the unidimensional nature of language proficiency. Knowing a language involved being able to use the language, and language proficiency is a combination of abilities; we conventionally talk about the four skills of listening, speaking, reading and writing (Heaton, 1975) and may divide each of them into various subskills. Other sources yield other analyses; Carroll for example suggests five broad skill types: oral and aural skills; graphic skills; language patterns (at and

below sentence level); discourse features (above sentence level); reference study and situation-handling skills. (Carroll, 1980: 22)

The elaboration of distinct language skills required that they be tested by distinctive means. In testing methodology, there is still an underlying competence, but this is inaccessible, and we can only get at it indirectly through different forms of surface manifestation. This observable behaviour comes in different forms, corresponding to different language skills; we therefore design different tests to test those different skills and from this period we retain the structure of the major dominant language proficiency tests today, such as IELTS (British Council) and the Cambridge main suite exams (UCLES) which contain a battery of sub-tests, including a direct or semi-direct oral test. While not comprising distinct sub-tests, the Five Star test provides an opportunity to test these skills, individually and in combination, through different task types within a single test event.

Concern over the objectivity of oral tests, and the need to improve the consistency of assessment of live oral tests, combined with the more systematic approach to syllabus specification, led to the development of marking keys incorporating those language features, mostly grammatical forms, which would be expected to have been mastered at each level of the test.

This model may be multi-dimensional, but two other assumptions remain: it is still linear - everyone is measured on the same continuum from beginner to advanced; and the traits are still stable, more or less. Inconsistencies between consecutive test results are seen as a poor reflection on the test, not as a good reflection of reality. Such test

techniques if they are consistently inconsistent should be discarded in favour of more consistently consistent ones. Even the assumption of multi-dimensionality was not universally agreed: as late as 1979, Oller was proposing a central language competence Unitary Competence Hypothesis (Oller, 1979).

Although not greatly concerned with the situational context of use, Oller was concerned with a linguistic context greater than the individual test item, and his enthusiastic espousal of integrative tests in contrast with the discrete-point atomistic paradigm ushered in the third of Spolsky's (1975) three phases, the psycholinguistic-sociolinguistic phase of testing. In particular, what he called global integrative tests, such as composition, cloze and dictation, are more than the sum of their parts and 'attempt to test a learner's capacity to use many bits [of knowledge of the language] all at the same time' (Oller, 1979: 37). However, because they are essentially indirect tests, Weir described these ten years later as 'the discredited holy grails of the psycholinguistic-sociolinguistic era' (Weir, 1990: 12). Oller also proposed a special sub-category of integrative tests drawing on the extra-linguistic context which he called pragmatic, and these are considered in 3.2.2 below.

2.7 Learning and acquisition

In a series of books and articles published in the USA from the late 1970s, Krashen (e.g. Krashen, 1981, 1982) developed a series of hypotheses that postulate, among other constructs, a distinction between acquisition and learning. *Acquisition* is the unconscious natural process by which we come to speak our mother tongue, and which

can also play a greater or lesser role in second or subsequent foreign languages. *Learning*, by contrast, is a conscious, formal activity. This contrast holds a strong echo of Chomsky's Language Acquisition Device (Chomsky, 1965), and also of Palmer's distinction between 'spontaneous' and 'studial' capacities (Palmer, 1922).

Krashen's interest in searching for a natural order of acquisition of elements of a second language, within a strong American academic tradition known as second language acquisition (SLA), culminated in collaboration with a language teacher Tracy Terrell to produce *The Natural Approach* (Krashen and Terrell, 1983). The similarity in name to the natural method, an alternative to the term Direct method mentioned above, was no coincidence; both sought to reflect ways we learn languages naturally, although there were importance differences.

In the Natural Approach there is an emphasis on exposure, or *input*, rather than practice; optimizing emotional preparedness for learning; a prolonged period of attention to what the language learners hear before they try to produce language; and a willingness to use written and other materials as a source of comprehensible input. The emphasis on the central role of comprehension in the Natural Approach links it to other comprehension-based approaches in language teaching (Richards and Rodgers, 1986: 129)

2.8 Humanistic perspectives

One of Krashen's specific hypotheses was the Affective Filter Hypothesis. This states that a learner's level of anxiety and attitude to the foreign language can promote or hinder his or her progress; a low affective filter will allow more input through. Research into attitude and motivation in language learning has been a continuing if

minor contributor to language learning theory, since the seminal work of Gardner and Lambert (1972).

A more direct concern for the learner as an individual was reflected in a number of specific applications of humanistic psychology from the mid-1970s, such as Silent Way, Community Language Learning and Suggestopedia. These are commonly if simplistically grouped together under the label 'humanistic perspectives'.

Silent Way is associated with Caleb Gattegno, a charismatic former teacher of mathematics who applied his insights in that field to language learning. While Cuisenaire rods and charts are well-known physical aids associated with Silent Way, there is a fully-developed rationale for which these aids are only the most visible and perhaps most trivial manifestation (Gattegno, 1976)

Community Language Learning (CLL) was developed by Charles Curran, a professor of psychology at Loyola University, Chicago. He employed counseling as a consistent metaphor for learning, and applied Rogerian counseling techniques, with CLL the specific application to language learning (Curran, 1976).

Suggestopedia was developed by the Bulgarian psychiatrist Georgi Lozanov. It aims to optimise learning and memory by the power of suggestion, and specifying a detailed attention to the learning environment involving music, rhythm, physical comfort, tone of voice, and methods of delivery (Lozanov, 1978). As well as, or perhaps because of, being the most dogmatic of the humanistic methods, suggestopedia has also aroused some of the strongest reactions (Scovel, 1979).

While it is difficult to be specific about the impact of these particular methods, they retain a niche on the syllabus of every ELT teacher education programme, and there is no doubt that they have influenced a generation of English language teachers, through their contribution to the progressively evolving views of other influential methodologists such as Stevick (1976, 1980, 1982). They “blend what the student feels, thinks and knows with what he is learning in the target language. Rather than self-denial being the acceptable way of life, self-actualization and self-esteem are the ideals the exercises pursue” (Moskowitz, 1978, quoted in Richards and Rodgers, 1986: 114)

There is a parallel between the 'learner as a whole person' of the humanistic methods and 'learner as real world communicator' of the communicative approach; the former has an internal focus within the individual and the latter an external one, but in both cases the emphasis is on treating the learner as a person of equal value with their own set of meanings and relationships. There is no catchphrase such as 'humanistic testing' but there is certainly a recognition that in order to test communication through a language rather than knowledge of it you must ask people to do real things that are meaningful for them. This is reflected in discussion in later chapters of communicative attributes such as authenticity, interaction and individualisation. Aspects of the design of the Five Star test and the arrangement of the physical setting of the test event (interviewer and candidate sitting side by side rather than the face-to-face confrontation of the typical oral interview) are consciously intended to make it as natural as possible (SDT no date) and reduce the lack of symmetry in the discourse of a typical oral interview test (Pollard, 1998a).

2.9 The communicative approach: introduction

A seminal article by Hymes (1970) criticised Chomsky's performance / competence contrast, not for being wrong, but for being irrelevant to practical needs. Hymes quotes Chomsky's delineation of the ground of linguistics:

Linguistics theory is concerned primarily with an ideal speaker-listener, in a completely homogeneous speech community, who knows its language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors ... in applying his knowledge of the language in actual performance (Chomsky, 1965:3).

Hymes comments "From the standpoint of the children we seek to understand and help such a statement may seem almost a declaration of irrelevance. All the difficulties that confront the children and ourselves seem swept from view" (Hymes, 1970, quoted in Brumfit and Johnson, 1979: 5) (he was originally speaking to a conference on 'Language Development Among Disadvantaged Children').

Hymes proposed a kind of competence that was much broader than Chomsky's, to include appropriateness and acceptability: "There are rules of use without which the rules of grammar would be useless" (page 15) and he is usually credited with coining the term 'communicative competence' for this. The essence is that the communicative effect of an utterance depends on much more than its mere grammaticality, and must allow for sociocultural and contextual factors.

In the mid-1970s, this construct filtered through to language teaching, and in conjunction with the situational approach led directly to the construction of non grammatically-based syllabuses:

The situational syllabus... is based upon *predictions of the situations in which the learner is likely to operate* through the foreign language... one can envisage planning the linguistic content according to the semantic demands of the learner... What is proposed, therefore, is that the first step in the creation of a syllabus should be consideration of the *content* of probable utterances and from this it will be possible to determine which forms of language will be most valuable to the learner. The result will be a *semantic* or *notional syllabus* (Wilkins quoted in Brumfit and Johnson, 1979: 83-84).

Wilkins' highly influential book was in fact called 'Notional Syllabuses' (1976). This essentially academic proposal was rapidly taken up by the English language teaching profession in particular, whose services were in growing demand for job-related rather than school-based or academic courses. There was a clear need for teaching, courses and learning materials which were directly related to adult learners' needs rather than the abstract grammatical syllabus that had dominated hitherto. The message was promulgated by the work being done by and through the Council of Europe and its attempt to establish a 'common core unit/credit system' among European languages (Wilkins, 1972). Political and social diplomacy also favoured the notional syllabus over the grammatical, in that situations of use could be described independently of any particular language, in theory at least, whereas a grammatical syllabus must be largely language-specific.

Between 1975 and 1985, the so-called 'communicative approach' (CA) or 'communicative language teaching' (CLT) revolutionised language teaching, and remains the dominant paradigm today. Richards and Rogers identify four underlying principles to CLT:

- Language is a system for the expression of meaning
- The primary function of language is for interaction and communication
- The structure of language reflects its functional and communicative uses
- The primary units of language are not merely its grammatical and structural features, but categories of functional and communicative meaning as exemplified in discourse (Richards and Rogers, 1986: 71)

They go on to characterise communicative activities based directly on these principles as involving real communication; carrying out meaningful tasks; and engaging the learner in meaningful and authentic language use. There is therefore a direct contrast with traditional and audiolingual practices of mechanical drills and pattern practice activities which serve no real communicative purpose.

Even the methodology for introducing the language to be used could be re-examined in the communicative terms. The systematic syllabus design of the oral or audiolingual approaches demanded that the teacher present the target language for the lesson before asking learners to try to use it, in controlled practice activities; Brumfit suggested a post-communicative procedure might require learners to communicate first of all as best they could with their available resources; this would throw up the language items they lacked but need for effective communication; which the teacher could then present and drill if necessary (Brumfit and Johnson, 1979:183). The task can be seen to be the focus of the lesson, rather than a pre-ordained syllabus of structure or lexis, foreshadowing the emergence of a methodology of task-based learning.

As well as a methodology of classroom activity, CLT was also realised through an approach to programme content, premised on the transparency and predictability of the situations or desired contexts of use of the individual learner. The 'strong' form of this was needs analysis, exemplified in its most extreme form by Munby's 'Communicative syllabus design' (Munby, 1978) which borrowed the terminology of automation ('communicative needs processor', 'linguistic encoder') to suggest that the process of needs analysis was both objective and finite. Here, and in other work directly influenced by it, such as 'Testing communicative performance' (Carroll, 1980), the word

'communicative' actually means 'based on a tightly-constrained analysis of the contexts of use needed by the learner'. Needs analysis is essentially identical to the specification of Target Language Use (TLU) now favoured by some authorities on testing (e.g. Bachman and Palmer, 1996) although the heuristic framework for the analysis may differ. The use of the word 'communicative' to imply specific domains of use for an individual remains an enduring ambiguity in all models of the communicative approach that include a sociolinguistic component.

In practice, needs analysis is most useful when it serves to confirm what intuition and common sense have already suggested. In a few cases, such as an airline hostess or a waiter (Munby's own examples), the contexts of use are highly predictable; in the great majority of cases, they are less so, and in neither case will needs analysis allow for the truly unpredictable utterance that is characteristic of genuine language use.

A unifying theme between CLT as methodology and CA as content specification has been the use of authentic materials and the re-creation of patterns of interaction and sequences of activity based on real life. This can become self-justifying, by suggesting that the use of authentic materials in an authentic context must necessarily be good practice. The nature of authenticity remains a live debate, and will be taken up below.

2.10 The communicative approach: testing

The term 'communicative' is in widespread use, in both teaching and testing contexts, and enjoys enough of a generally understood core meaning to be useful. However, part

of the difficulty of describing the communicative approach to testing lies in the shades of meanings used by different authors, without always making their assumptions explicit. The communicative approaches can be included in the third of Spolsky's (1975) three phases, the psycholinguistic-sociolinguistic phase to testing.

The evolution of the communicative methodology in language teaching required a parallel development in testing methods, with an emphasis on eliciting and demonstrating 'language in use' in authentic contexts rather than 'language as knowledge' in conventional pencil-and-paper tests (Carroll, 1980).

Three major implications for language testing were firstly, that tests should be based around meaningful tasks rather than unrelated items (Brown, 1994). These tasks should be based in a context with a recognisable parallel with the real world; associated criteria implied here are authenticity, purpose and unpredictability of communication (Morrow, 1979:149-150)

Secondly, to be authentic, test techniques should be interactive and allow the production and comprehension of language at the level of discourse, i.e. operating with language samples larger than the single sentence or short utterance (Carroll, 1980). Most language use in the real world involves more than a single utterance or sentence and is based on two-way communication, whether spoken or written; it is artificial to try to separate out the listening and speaking components of an interaction, when what someone says depends largely on what is said to them.

A third broad aspect of communicative methodology has been the identification of strategic competence, as distinct from purely linguistic competence, as a focus for both teaching and testing (Canale and Swain, 1980). This construct covers interaction skills and non-verbal communication (Milanovic et al., 1996).

Weir (1990: 10-14) summarises the distinguishing features of communicative language tests. He describes the need for:

- a) the identification of test purpose and matching of tests and tasks with target language use, with the implication that 'no one solution can accommodate the wide variety of possible test scenarios', i.e. there cannot be an 'all-purpose' communicative test
- b) contextualisation, including an 'integrative approach to assessment'
- c) authenticity of tasks and genuineness of texts (see 2.9.5 below)
- d) reflection of the interactive nature of normal spoken discourse
- e) tasks conducted under normal time constraints and with an element of unpredictability in oral interaction
- f) direct testing requiring performance in integrated skills from the candidate

These principles have had a profound effect on language testing in the last 10 years, and have been implemented in a number of more recent language tests and examinations, with varying degrees of success (Alderson et al., 1995). This is an ideal profile which is necessarily compromised in the process of implementation into actual test instruments, but these remain distinguishing features of the communicative approach to testing, and are reflected to a greater or lesser extent in the design of most tests today. The Five Star

test is however rare in the extent to which tasks require the use of integrated skills, rather than testing them separately as most current tests still do.

Problem areas for communicative language testing

The problems language testers were being presented with by the communicative approach were recognised at an early stage.

For many years, linguistic criteria were the only ones considered in language testing. A person's ability to communicate was assumed to be related directly to his ability to control the linguistic elements of the language Later fluency was added to the list, even though it was not at all certain that everyone agreed on what it meant. Recently, a rising interest in communicative competence has forced us to examine more closely what else besides linguistic facility contributes to effective communication ... Unfortunately, communicative competence has come to mean many things to many people, and it is not a term that is unambiguously understood among language teachers. But certainly a sensitivity to appropriateness of language and an understanding of nonverbal paralinguistic signals are important. Unfortunately, these additional features pose very difficult problems for testing (Madsen and Jones, 1981: 21)

As well as the traditional linguistic skills, the communicative and situational approaches introduced extra-linguistic factors, such as social context, pragmatics, interaction skills, strategic competence, which combine together to form communicative competence.

This clearly needed to be reflected in the testing context, but raised the problems of how to test the language systems in combination and how to allow for the extra-linguistic factors. Davies pinpointed this as a problem for communicative testing, where it was not a problem for communicative teaching:

Language tests are capable of handling language elements, testing for control of elements in the grammar, the phonological systems etc. This is the well tried discrete point approach. But the very nature of language systems is that they combine, grammar with semantics, etc. Now a test of the overall language systems (i.e. of all the systems in combination) may be no more than a set of tests of the systems... What is clear is that linguistic descriptions do not provide any compellingly satisfactory account of system combination. Therefore what passes for a test of language systems is either a disguised grammar test or a confusing mixture, a kind of countable shotgun approach to combination ... the point is that there is no rationale for designing a language test to test language systems in combination because there exists no algorithm which will tell us how language systems combine. Davies (1990: 185)

This helps to explain why current tests that reflect many of the communicative criteria listed above still do not test integrated skills within tasks. A counter-argument sometimes put forward in defence of discrete-point tests is on the grounds that the skills they test still need to be taught on an analytic basis and should therefore be tested in the same way: "Rhetoric touting performance assessments because they eschew decomposed skills and decontextualised tasks is ... misguided, in that component skills and abstract problems have a legitimate place in pedagogy" (Messick, 1994: 13) but there is a risk of circularity here. Messick is in effect justifying testing practice on the grounds that it reflects classroom practice, when in fact the classroom teaching methods may themselves be strongly influenced by washback from tests. Davies continues:

The area of most difficulty for language testing is in the attempt to develop true performance tests, a good example of which is communicative language testing. The impetus to develop communicative language tests comes from the communicative language teaching movement The problems that arise ... are precisely that language testing needs to be clear, in this case explicit, about what is under test. Ironically there is no such constraint on communicative language teaching ... [which] can tolerate the central uncertainty of ...tasks and not demand explicitness. But that is precisely what tests, as tests, must do. In testing this is desirable both for the language input and for the measurement output.' (Davies, 1990: 185-187)

2.11 British and American approaches to communicative testing

It is useful, if simplistic, to contrast two approaches to the development of communicative testing since the 1970s and to stereotype these as the British/European approach and the North American approach (these are crude but convenient labels, with exceptions in both camps).

In the British/European school, the impetus for the development of communicative testing came from the work in the early and mid 1970s on the notional/functional

syllabus associated with Wilkins, Van Ek and the Council of Europe. This work was applied in the late 1970s to the operational problems posed by communicative testing, leading to Morrow's seminal work *Techniques of evaluation of a notional syllabus*, (1977) a booklet produced for the Royal Society of Arts (RSA) specifically to provide a practical basis for the development of the RSA's new test of communicative competence, the Communicative Use of English as a Foreign Language (CUEFL).

The RSA Examinations Board was subsequently merged with the University of Cambridge Local Examinations Syndicate (UCLES), and over the last 15 years much of the RSA's ground-breaking work on communicative testing has been taken up and applied to UCLES's main suite of EFL exams. Interestingly, one of the four skills tested by CUEFL was originally called 'oral interaction', although it has now reverted to the more common name of 'speaking'. A widely used model for identifying and classifying target language was Munby's Communicative Syllabus Design (1978) mentioned above, which presented a deterministic needs analysis model to predict the actual language needed by specific learners.

Linguistically, this approach drew on the work of Halliday, Sinclair and ultimately Firth, who saw language as a social system for the conveyance of meaning, inseparable from its context of use, to be studied at the surface level of observation of real data, in order to identify patterns and regularities.

The American testing tradition

In the north American school, the seminal theoretical model for communicative testing was put forward by Canale and Swain (1980, 1981), and builds on Morrow's work. They proposed three main components to communicative competence: grammatical competence, sociolinguistic competence and strategic competence, but followed Chomsky's deep/surface distinction by suggesting that communicative competence is observable only indirectly through its manifestation at surface level in communicative performance.

Tests of practical language skills are sometimes called performance tests, and may be underpinned by a theory of communicative competence to provide a theoretical rationale. Performance tests are expected to reflect the use of language in the real world, and the labels 'direct' and 'authentic' are often used for this; in some contexts, 'communicative' is also used as another synonym for this criterion of lifelikeness. Authenticity is almost as elusive and problematic a construct as communicativeness, and its use as a criterion for validity is taken up in 2.9.5 and 3.3.1 below. For McNamara (1996), it is the activity of the rater, rather than the behaviour of the test participant, that is the critical characteristic of a performance test: "performance necessarily involves subjective judgment" (1996: 117)

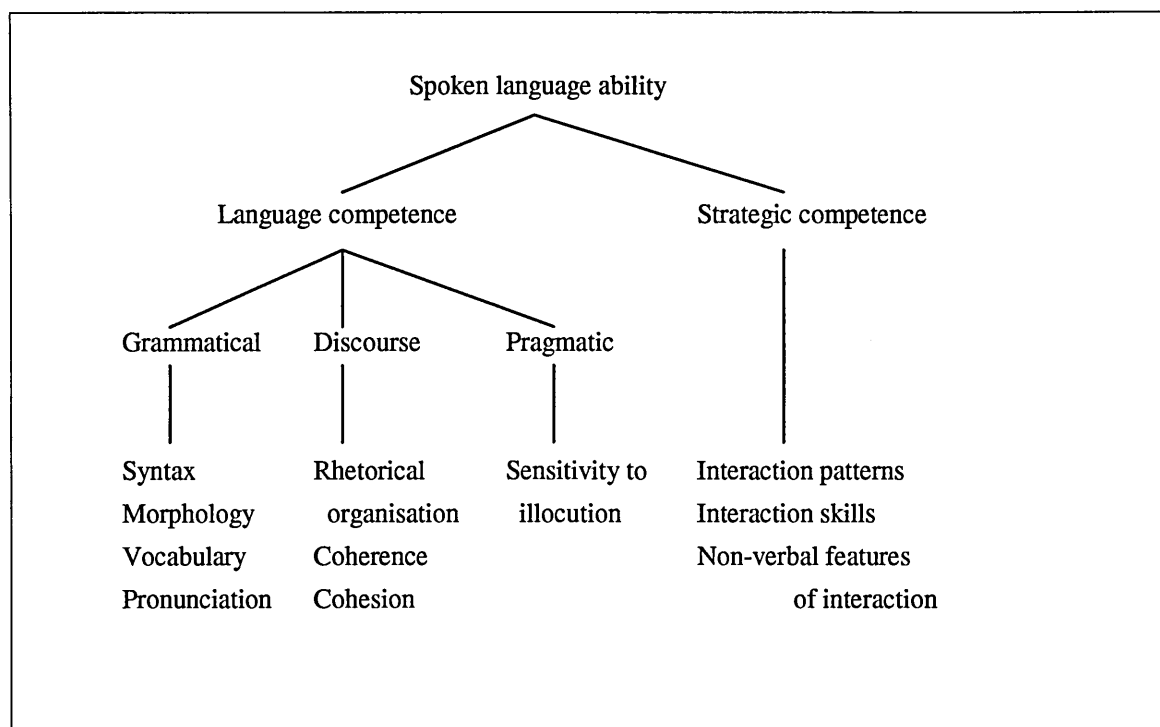
This approach by contrast drew directly on Chomsky's perception of linguistics as a theoretical branch of cognitive psychology seeking to identify the underlying universal rules of language competence to account for isolated examples of data generated by the intuition of the native speaker.

For Canale and Swain, grammatical competence 'will be understood to include knowledge of lexical items and of rules of morphology, syntax, sentence-grammar semantics, and phonology' but without specifying or preferring any particular school of grammatical analysis. Sociolinguistic competence

is made up of two sets of rules: sociocultural rules of use and rules of discourse. Knowledge of these rules will be crucial in interpreting utterances for social meaning, particularly when there is a low level of transparency between the literal meaning of an utterance and the speaker's intention. Strategic competence will be made up of verbal and non-verbal communication strategies that may be called into action to compensate for breakdowns in communication due to performance variables or to insufficient competence. (Canale and Swain, 1980:28-31)

Within each component, there is envisaged a 'subcomponent of probability rules of occurrence', which will attempt to characterize 'the knowledge of relative frequencies of occurrence that a native speaker has' with respect to each competence, e.g. the probable sequences of words in an utterance as an example of grammatical competence.

The whole Canale and Swain model has been adapted by the Cambridge Assessment of Spoken English project, as illustrated in Figure 1 (Milanovic et al., 1996, in Cumming and Berwick, 1996; Saville and Hargreaves, 1999).

Figure 1**Model of spoken language ability used in the CASE project**

Examples given by Canale and Swain of communication strategies as manifestations of the underlying strategic competence are the ability to paraphrase or circumlocute when the exact word or phrase is unknown or temporarily forgotten, and coping strategies, such as role-playing. Well-known taxonomies of such strategies have been prepared by Tarone (1980) based on an interactional approach and Faerch and Kasper (1983) taking a more psycholinguistic perspective. Summary taxonomies proposed specifically for the purpose of analysing interaction in oral tests are put forward in two separate accounts in Young and He (1998), by Yoshida-Morise and Katona.

Three major common categories are reduction (or avoidance) strategies, for example, remaining silent or abandoning a message; achievement (or compensatory) strategies, such as circumlocution, approximation or translation; and stalling or time-gaining

strategies, such as using filler words and hesitation devices. Participants' differential use of such communication strategies is influenced by factors such as personality, the learning environment, task demands and proficiency level (Yoshida-Morise in Young and He, 1998: 215).

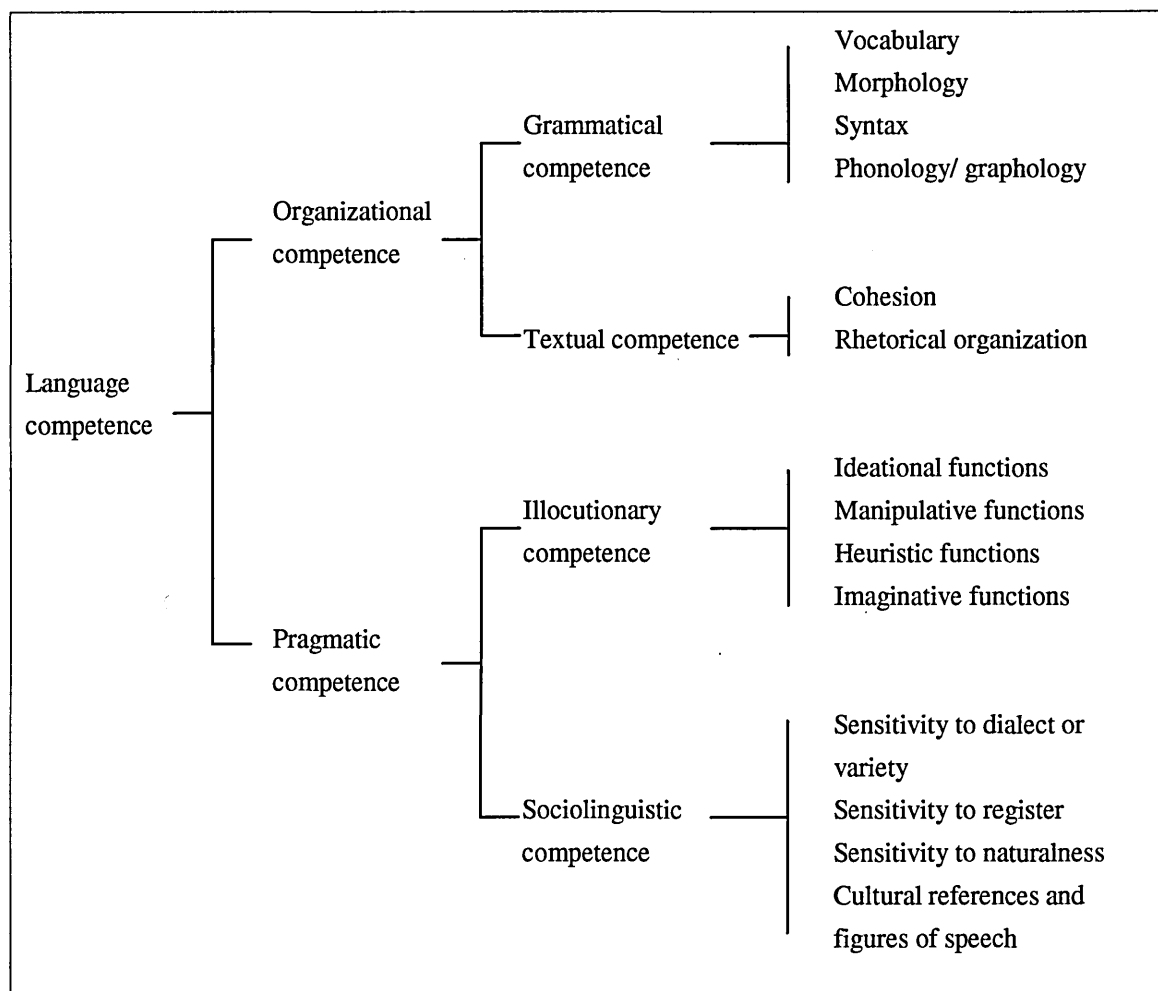
The exploration of sociolinguistic competence has been given new impetus in the last few years by the growing interest in the new inter-disciplinary field of cross-cultural studies. Research into oral test interaction from this perspective suggests that misperceptions of the illocutionary force of interview questions may lead candidates to fail to display the language proficiency of which they are capable. Ross (in Young and He, 1998:333) uses the term 'misplaced underelaborations' to describe situations in which the candidate fails to decode an interview question as an invitation to speak about a topic, and may provide a minimal response. In the specific case of the Japanese culture which Ross considers, the candidate may also be operating to cultural norms that seek to avoid verbosity, to be reluctant to discuss matters in the personal domain and to be unwilling to air personal opinions spontaneously. These socio/cultural rules create the risk of violating the maxim of quantity, from the interviewer's perspective, by generating an insufficient language sample, and so giving the impression of a candidate who is not fully committed to the interaction or who has limited language proficiency.

The area of sociolinguistic or strategic competence figures in many testing and scoring systems which have not adopted the communicative competence model wholesale. The University of Michigan's Examination for the Certificate of Proficiency in English, for example, has a direct test of spoken English which includes 'functional language use/sociolinguistic proficiency' among the list of salient features that examiners are to

look out for, which in turn includes sub-components of interactional facility, sensitivity to cultural referents, and others (ELI, 1999).

Building on Canale and Swain's work, Bachman developed a more elaborate set of components of communicative language ability in 1990, shown in Figure 2.

Figure 2 **Bachman's 1990 components of language competence model**



The components of language competence in this model are distinct from strategic competence, which is 'the mental capacity for implementing the components of

language competence in contextualised communicative language use' (Bachman, 1990: 84).

Problems with componential models of communicative competence

Conscious of the growing complexity of his analysis, Bachman points out one of the dangers of this kind of branching model:

This 'tree' diagram is intended as a visual metaphor and not as a theoretical model, and as with any metaphor, it captures certain features at the expense of others. In this case, the diagram represents the hierarchical relationships among the components of language competence, at the expense of making them appear as if they are separate and independent of each other. However, in language use these components all interact with each other and with features of the language use situation. Indeed, it is this very interaction between the various competencies and the language use context that characterizes communicative language use. (Bachman, 1990:87)

Another risk that such models run is that of reductionism. The diagram has four layers already, and particularly in linguistic areas that have been well-researched, such as syntax and morphology, it would be quite possible to add further layers listing more components of the model without necessarily saying how they combine or interact. In other words, it is descriptive rather than explanatory. Bachman concludes, without apparent irony, that "attempts to empirically validate these various components have not been conclusive." (1990: 86)

A slightly revised version of this model is presented in Bachman and Palmer (1996), with some changes of terminology which make it appear less remotely theoretical more accessible to practical exploitation by language testers.

Firstly, 'metacognitive strategies' replace 'strategic competence' and its three major subcomponents are

- a) Goal setting - identifying test tasks and deciding which to tackle
- b) Assessment - assessing the demands of the task and how one's own knowledge matches those demands for a correct and appropriate response
- c) Planning - deciding how to use one's knowledge

The emphasis here seems to be more directly concerned with evaluating the immediate context of use and responding appropriately.

Secondly, the word 'knowledge' replaces the word 'competence' throughout the language component model shown in Figure 2, so that for example, 'pragmatic competence' becomes 'pragmatic knowledge'. There is no explanation for this systematic change; one might speculate that 'competence' is a more loaded word in the North American approach with the Chomskyan implication of a deep level construct that is beyond direct scrutiny, whereas 'knowledge' is more amenable to measurement.

For practical purposes however, the complexity of the proposed communicative model still makes it virtually impossible to replicate it for use in a practical and economical way in language testing programmes. A current perspective on the value of such complex componential models for practical test construction might be summarised as a 'check-list' approach:

the value of Bachman's model is in its weak version, which provides researchers with a useful organizing structure within which systematic language testing investigations can be established... Assessments that are typically constructed in a given context will not include all the features depicted in [Bachman's model]; only those aspects highlighted by the variables operating in that context will be salient (Chalhoub-Deville, 1997:8).

In an analysis of questions and answers in an oral proficiency interview, He (in Young and He, 1998: 101) found discourse competence to be inseparable from grammatical competence, because of the difficulty of identifying the reasons for features such as

pauses in speech, reluctance to elaborate and the problematic use of discourse markers such as 'yeah'.

Davies' 'countable shotgun' approach describes most actual test instruments, where requirements of time, physical resources, and staffing for construction, administration and marking would make it impractical and very costly to apply such a model fully. Everyday test applications diverge from models derived from testing research at the point where accurate information on the language ability of individuals becomes an expensive commodity. (Hughes, 1990)

2.13 Work sample and cognitive approaches to assessment

The contrast above between the British/European and the North American approaches is echoed in the distinction made by McNamara between two traditions of second language performance assessment, which he calls the work sample and the cognitive approaches (McNamara 1996). The work sample approach developed from the practical needs of personnel selection and academic admission, and is 'resolutely pragmatic and atheoretical... behaviour-based and sociolinguistic in orientation' (1996:25). It is premised, as the label implies, on the assumption that direct sampling of the target behaviour is both feasible and a viable basis for assessment. The performance of the candidate is the target of the assessment (Messick, 1994)

The cognitive approach in contrast has a more theoretical psycholinguistic underpinning, 'focusing less on the verisimilitude of performances and more on what

they reveal about underlying ability and knowledge' (McNamara, 1996: 25). The motivation for its development was the need to bring performance testing into line with the communicative approach in teaching, an effect that is sometimes known as 'bow-wave' (in contrast to the more common 'backwash' impact of test design on teaching methodology). Here, the performance is the vehicle rather than the target of the assessment, in the belief that what it reveals is of more interest than the performance itself.

McNamara's work sample vs. cognitive distinction draws explicitly on a contrast between task-driven performance assessment and construct-driven performance assessment made by Messick in 1994. In the former, the performance itself is the target of the assessment and the focus for validation is on judgement of the quality of the performance. Replicability and generalizability of performance are not at issue, but because of the focus on performance, inferences cannot be made about underlying concepts such as knowledge or skills. As Weir says, 'Strictly speaking, a performance test is one which samples behaviour in a single setting with no intention of generalising beyond that setting' (Weir, 1990: 7) but accepts that in practice a distinction between testing performance and testing competence is impossible to maintain.

In construct-driven performance assessment, on the other hand, the performance is not so much the target as the vehicle for the assessment of competences or other constructs. Replicability and generalizability are important 'because the consistency or variability of the performances contributes to score meaning, as does generalizability from the sample of observed tasks to the universe of tasks relevant to the knowledge or skill

domain at issue' (Messick, 1994: 14-15) which constrains the interpretation of test scores. Table 1 summarises this contrast.

Table 1 **Task-driven and construct-driven performance assessment**

In task-driven performance assessment:	In construct-driven performance assessment:
<ul style="list-style-type: none"> ▪ the performance itself is the target of the assessment ▪ Validation focuses on judgement of the quality of the performance; inferences cannot be made about underlying concepts such as knowledge or skills ▪ Replicability and generalizability of performance are not at issue; ▪ ... but the transparency and meaningfulness of the task become more important 	<ul style="list-style-type: none"> ▪ Performance is the vehicle for the assessment of competences or other constructs ▪ Validation focuses on what performance reveals about underlying constructs such as components or skills ▪ Replicability and generalizability contribute to score meaning; ▪ ... and the meaning of the construct is tied to and limited by the range of tasks and situations it can be generalised to.

Messick's main concern in this article is that performance tests should be subject to the same stringent validity criteria as any other tests, and his incorporation of authenticity and directness within the framework of construct validity is considered in section 3.3.1 below. However, he notes an important specialised validity criterion referred to as *transparency* (Frederiksen and Collins, 1989) or *meaningfulness* (Linn et al., 1991), and uses the term *credibility* himself, described as 'the extent to which the criterion situation is faithfully simulated by the test' when 'the problems and tasks posed [are] meaningful to the students.' (Messick, 1994: 16-17).

He mentions some possible techniques to improve the transparency and meaningfulness of tasks, such as using subjective scoring or providing more realistic item contexts. This

emphasis on consideration of the students' (and teachers') response to task types and scoring systems is an example of consequential validity, and allows for a positive washback effect to contribute to overall validity. However, his use of the term *credibility* seems very close to what is elsewhere called face validity, but became discredited for lack of an empirical basis (Stevenson, 1981). Although calling for performance tests to be subject to the same requirement for empirical evidence of validity as other tests, Messick does not describe how evidence for credibility can be measured empirically.

The issue of generalizability (or 'extrapolation' in Morrow, 1979; Weir, 1990) is related to the question of how target language use is sampled and how tasks can be selected for a test on some kind of systematic basis. One solution offered is to identify through an analytic approach to the target language what are called 'enabling skills' which are evidenced in performance in particular tasks but are also claimed to transfer or extrapolate to other tasks which are not directly tested. As Morrow acknowledges, 'the status of these enabling skills vis-à-vis competence: performance is interesting' (1979: 152) because they appear to bridge the gap between underlying competence and observable performance, yet cannot be satisfactorily treated either as deep-level constructs or measurable variables.

Another contrast that partially overlaps with the British work sample vs. North American cognitive approaches is that between testing as a complementary activity to teaching and as a research-based sub-discipline of applied linguistics. One of the foremost language testers in America, for example, recently wrote 'developments in language testing research ... have brought language testers into close contact and affinity

with their fellow applied linguists, as well as with specialists in measurement and technology' (Bachman, 2000: 1) yet teachers do not get a mention.

2.14 The communicative approach: individualisation, task-based learning and authenticity

Three important sub-themes that have emerged from communicative language teaching over the last ten years are individualisation, task-based learning and authenticity.

Individualisation and learner-centredness

Drawing directly on the needs analysis model of the communicative approach is a specific concern for each student as a distinct individual with distinct needs, styles and strategies. The implicit assumption of almost all research in second language acquisition (SLA) has been “towards establishing how learners are *similar*, and what processes of learning are *universal*” (Skehan, 1989:1) and this assumption is carried over to teaching methods and materials. A learner-centred approach turns this assumption around, and takes as a starting point that there are in fact differences between learners; as individuals, they have an internal syllabus which may be better satisfied by a task-based process approach rather than an external product-based syllabus. The first implication of this is a shift in focus from teaching (what the teacher does to a class of learners) to learning (what an individual does in a class, and elsewhere). A second implication is that individuals may differ significantly not just in their language abilities but in their cognitive styles and preferred learning strategies (Skehan, 1989).

Task-based learning

The second methodology derived from CLT that has acquired an identity of its own is task-based learning. A contrast associated with this is that between process and product, where the value of a learning activity may not lie in the outcome (for example, a learner completing a task properly or answering certain questions correctly) but rather in the processes gone through in reaching that outcome. This reflects the substantial part of human communication that is not purely concerned with the transmission of factual content, and as the processes at work are for the most part hidden, it encourages individualisation to operate in the classroom.

It is difficult to be precise about exactly what a task is, and why other teaching and testing activities may not be considered task-based. In a recent paper, Bruton considers various definitions, including one that emphasizes the meaningfulness of content: "... task: a goal-oriented communicative activity with a specific outcome, where the emphasis is on exchanging meanings not producing specific language forms" (Willis, 1996: 36 quoted in Bruton, 1999). However, there is a risk of circularity here, if the communicative methodology is exemplified by task-based learning, and tasks are in turn defined as emphasizing the communication of meaning rather than language form.

Bruton's own definition is more complex:

A task is an activity with an identifiable goal set by oneself or another, which might suppose some difficulty in its achievement, requiring special effort and usually assuming a limited duration of time. The means of achieving the goal might be pre-planned, and in some cases may be completely routine, but the outcomes of a task may not always be predictable. (Bruton, 1999:2)

This is contrasted with a project, which is "... an extended piece of work on a particular topic where the content and the presentation are determined principally by the learners" (Hutchinson, 1991: 10 quoted in Bruton, 1999)

Individualisation and task-based learning: testing

These recent trends have further complicated the near-impossibility of genuinely communicative testing. How can tests be individualised and yet comparable? How can you validate a test when no two administrations are the same? Adaptiveness offers the possibility of individualised testing, but raises the issue of comparability of language samples, a problem for conventional validation methods. If people learn in significantly different ways, they may be good learners and bad test-takers or vice-versa; to be consistent, differences in teaching methodology must be reflected in testing.

The Five Star test allows the interviewer to build in a high level of individualisation through the interaction with the candidate. It substantially reflects a task-based approach through the assessment in many tasks (but not all) of the language sample generated rather than the enumeration of right or wrong answers.

The introduction to a collection of articles specifically focusing on this issue *Individualising the assessment of language abilities* (De Jong and Stevenson, 1990) identified three factors influencing the development of individualised testing. The first is a greater awareness of the complexity of language and of the diversity of acceptable models; the second is the availability of information and communication technology, constantly increasing in power and decreasing in cost; and the third is public attention to educational accountability, questioning the fairness of norm-based judgements.

From this perspective, individualisation does not necessarily mean a different test for every person; it is rather a contrast with what they term 'broad-spectrum testing'. In the same volume, Spolsky describes norm-referenced testing as "potentially dangerous to the health of those who take it" (Spolsky, 1990: 12) Thus, any movement towards designing tests for specific populations or to meet specific needs will represent individualisation. Similarly, any provision for reporting the results that is more informative than a single summary score, for example by use of rating scales or performance descriptors, could be considered to be individualised.

To the extent that broad-spectrum testing tends to provide less information the more an individual 'deviates' from the norm, individualized testing is in one way simply fulfilling what test developers have long recommended, but have never been able to acquire. This is that test users tailor the tests, the uses of the test, norms and interpretive guidelines to their own populations and educational needs. (de Jong and Stevenson, 1990: xiv)

A framework for describing language tasks is proposed by Bachman and Palmer, with the dual purpose of describing both real-world and test tasks (1996: 47). Real-world language is labeled Target Language Use or TLU. The aim of the framework is to enable description of the TLU tasks and of different test tasks, and therefore a common framework for comparing the characteristics of TLU and test tasks to assess their authenticity. The framework is in the form of a checklist, with five major aspects: setting, test rubric, input, expected response and the relationship between input and response. Each of these aspects is broken down into further component characteristics, so for example, 'rubric characteristics' includes the task instructions, number and structure of the tasks, timing, and scoring method, and 'response characteristics' include 'Language of expected response' analysed by the components of language competence model (Bachman, 1990; Bachman and Palmer, 1996) outlined in section 2.9.2 above.

The discussion in the previous section about how to define a task raises the question of how to distinguish between a test task and a typical test item, for example, a multiple choice or true/false question. Bachman and Palmer's framework takes the simple solution by not attempting to define 'task' in such a way as to exclude any language-based test activity. Thus for them a multiple choice item is a test task, but one which differs significantly from a composition or oral interview in its characteristics analysed according to their framework. There is no *a priori* judgemental evaluation from a communicative perspective, as suggested by Bruton's (1999) definition quoted above; the evaluation of the usefulness of a task is made by comparison against the identified target language use. It is in effect what McNamara would call the work sample approach, using a needs analysis as the ultimate yardstick. This seems superficially more objective, but it is a deterministic model that assumes a degree of predictability and precision in the description of TLU tasks that is often unrealistic.

A further objection to the needs analysis model is that while specific language needs may indeed be determinable, their use in real life is not divorced from 'general' language use. Van Lier felt that 'researchers in this area are becoming increasingly aware of the fact that professional interaction is embedded in conversational interaction ... and that success in the former depends to a large degree on the interactants' skills in the latter' (van Lier, 1989:500)

If communicative tasks are to be distinguished from ordinary test items, common sense might suggest that a task took longer to complete than a simple test question, and involved an element of discussion, transfer of information, or negotiation of meanings; in other words, a focus on communication that is absent from discrete-item tests. A

more principled criterion might be based on the process/product distinction; that the purpose of a discrete-item question is solely to determine whether a candidate has the knowledge to produce the correct answer or not, however they arrive at it, whereas the concern of task-based assessment is the language that is generated in arriving at the outcome. The outcome itself may have no single correct answer and will typically be irrelevant to the assessment. In this case, the distinction is linked intimately to the method of assessment, in that a task-based approach to testing can be characterised as requiring a subjective marking scheme which can cope with the complexity of a 'process' language sample longer than the isolated sentence or utterance.

The opportunity to make decisions in communicating was one of van Lier's three preconditions for evaluating oral proficiency. The choices to be made can include when to speak, for how long, and about what. A second precondition is goal-relatedness, requiring that 'linguistic ... tools used in the interaction are not the only yardstick for the evaluation of quality' (van Lier, 1989: 494). Both of these, decision-making opportunities and goal relatedness, might figure in a list of criteria for defining communicative tasks.

Authenticity in language testing

The requirement for authenticity in the communicative approach, and so in communicative language testing, has also sparked a lively debate. In the 1970s, Widdowson contrasted 'genuineness', which he saw as a measure of lifelikeness of a text in itself, with 'authenticity', which depended on the appropriateness of response of the reader or audience (Widdowson, 1979). The latter therefore overlapped with the

criterion of interaction. It was at least possible to establish genuineness; if a tester took an excerpt from a newspaper or radio broadcast and used it without editing, it could be called genuine. This created an oversimplistic dichotomy between genuine texts, which were seen as intrinsically good, and those which were edited or written specifically for teaching or testing purposes, which were intrinsically bad. As a result a lot of inappropriate materials were used out of context (Lewkowicz, 2000).

A more complex view of authenticity was taken by Bachman, who contrasted situational and interactional authenticity (Bachman, 1990). Situational authenticity was the extent to which a test task matched the target language use (TLU or real-life) task. Interactional authenticity, as in Widdowson's distinction, depended on the outcome of the activity, when the test-taker engaged with the task. Subsequently, he separated the two, defining authenticity as 'the degree of correspondence of the characteristics of a given language task to the features of a TLU task' (Bachman and Palmer, 1996: 23) while what had previously been called 'interactional authenticity' became 'interactiveness'. They proposed the checklist for matching the test tasks against the TLU task characteristics described in the previous section.

Like the needs analysis model, this framework is based on the underlying assumption that it is in fact possible to determine with confidence and accuracy the characteristics of real-life situations of language use. That this is not always possible can be seen as a good reflection of the authentic unpredictability of real-life language use, rather than a poor reflection of a particular test task. The Bachman and Palmer framework is only a checklist, and does not account for the combination or degree of centrality of the different criteria.

They also suggest that stakeholders' perceptions of authenticity vary. In an experiment aimed at eliciting which test attributes test takers identify as positive and negative, Lewkowicz (2000) gave a group of students a multiple-choice pen-and-paper test and an EAP (English for academic purposes) test, the latter having a reasonable claim to authenticity by virtue of similarity to real-life academic study tasks. Only 12% of participants identified authenticity as a positive feature of the EAP test, and Lewkowicz concluded that it was not an important test feature for most test-takers, and that 'there may be a mismatch between the importance accorded to authenticity by language testing experts and other stakeholders in the testing process'. However, since she relied on students' perceptions to generate the positive and negative features, and did not prompt or suggest any features to rate or prioritise, it may be that it simply did not occur to them to suggest features which they might otherwise have agreed with.

An alternative approach to authenticity has been to make reference to Searle's concept of speech acts (Searle, 1969) and the conversational maxims proposed by Grice (1975). Asking a question to which you already know the answer, and do not sincerely want the information requested fails to meet Searle's conditions for a real question. Searle recognised the problem and made a distinction between 'real questions' and 'exam questions' (1969:65); in 'exam questions', the speaker doesn't necessarily want to know the answer, she wants to know if the hearer knows. The illocutionary force is a request, rather than a question, to display certain knowledge or skills, and the onus is on the candidate in a test to recognise this and knowing the rules of the game to cooperate with the examiner in a willing suspension of normal speech conditions. Spolsky (1990) presents this as a social contract: 'an understanding on the part of the person taking the

test that the performance is necessary and relevant to the task' and concludes that 'the most we can then ask for is an authentic test; not authentic communication' (1990: 11).

However, in direct tests of oral performance, there can be shades of transgression of these speech act conditions, and these are discussed further in 7.2.6.

2.15 Discussion and summary

This chapter has put current issues in language testing in the context of its historical development, which has closely followed the evolution of language teaching. This in turn has been shaped by developments in other fields and by political and demographic changes.

The continuing rise in demand for English as an international language has focused attention on the testing of English above all other foreign languages, and in doing so has raised new issues. It has created a tension between global and local testing, between the desire to test proficiency as a means of global communication, to be used anytime, anywhere with participants of any background, and the communicative requirement to match tests to target language use. It challenges our assumptions about the extent to which communication in a natural language such as English owes to the cultural assumptions of its geographical origins rather than being 'culture-free' and the recent development of the whole field of cross-cultural communication reflects this concern. The Five Star test attempts explicitly to validate international English in a local context,

where the participants and other parameters of the interaction can be largely specified but where the involvement of native speaker participants and values cannot be assumed.

This chapter has looked at the methodology of teaching and testing in general, and has identified a long-term shift from focusing on the form of language as a closed system to the function of language as a means of everyday communication. The use of discrete-point tests based on an atomistic approach to language analysis is now seen as a fruitlessly reductionist endeavour and has largely given way to more integrative activities that do not pretend to reveal an underlying taxonomy of constituent parts, but rather aim to capture the more complex reality of 'language in use'. More recent models of language proficiency incorporate higher-level components and allow for the contribution of strategic, interactional, and even non-linguistic skills to the success of communicative competence, but the danger of reductionism remains. There are no compellingly satisfactory models to describe how and in what proportions these components combine; these models remain descriptive rather than explanatory, and there is no prospect of arriving even at a level of description that can claim to be unique. There is therefore no accepted theoretical framework at present on which to build language tests.

There is, however, considerable agreement on some of the desirable features that between them characterise the communicative approach. These are picked up in chapter four to describe the Five Star test. A further recurrent problem for the description of test practice is the diversity of shades of meaning of key terms such as 'authentic', 'interaction' and 'task-based' and indeed the word 'communicative' itself can have fundamentally different connotations when used in phrases such as 'communicative

competence', 'communicative performance' and 'communicative task'. An epistemological problem that is unique to research in language learning, teaching and testing is that language is both the object of study and the means of description. In Stevenson's metaphor, quoted in section 3.2.3 below, we are trying to measure something with tools made of what we are trying to measure, and the problem is to distinguish the tool from the matter.

Language testing specifically has also been strongly influenced by the field of psychological measurement, and this is particularly evident in the theory and practice of test validity. Chapter three now focuses on test validity and in particular looks at the distinctive nature of testing spoken language, as opposed to the testing of writing or testing of language through other forms of response such as multiple choice.

Chapter 3 Literature review 2: Validity in language testing; testing spoken language; adaptive testing and item response theory

- 3.0 Introduction
- 3.1 General introduction to validity and validation in psychometric testing
- 3.2 Validity in language testing
 - 3.2.1 Validity as a single entity
 - 3.2.2 Validation of general language proficiency
 - 3.2.3 Construct validity as an empirical construct
 - 3.2.4 Validation in the rationalist tradition
 - 3.2.5 Content validity
 - 3.2.6 Validity as a unified concept
- 3.3 Validity issues in oral testing
 - 3.3.1 Direct and authentic
 - 3.3.2 Interaction
 - 3.3.3 Face validity
 - 3.3.4 The empirical validation of oral tests
- 3.4 Adaptive testing
- 3.5 Item response theory (IRT)
- 3.6 Summary

3.0 Introduction

Building on the general background to language testing outlined in chapter two, this chapter focuses on issues of validity in language testing. It looks at the way the concept

of validity has evolved through the different sub-categories of validity that have been envisaged and how they have been measured. An introduction to validity in psychometric testing is followed by an overview of the historical development of validity in the language testing literature. Subsequent sections focus on three validation issues of particular relevance to the present research; the validation of spoken language tests, adaptive testing and the contrast between classical and item-response theory statistical techniques.

3.1 General introduction to validity and validation in psychometric testing

Validity is an essential attribute of every kind of test. Broadly, it indicates simply whether a test is in fact measuring what it purports to measure. At a very general level, it may be grouped with other desirable test attributes, such as efficiency (or economy) and reliability, but even reliability is sometimes seen as a specific sub-type of validity.

There is an extensive literature on validity, both in psychometric testing in general and language testing in particular, and as soon as discussion moves from the general level of concept to the operationalisation and measurement of validity, its coherence as a single construct disappears and a variety of sub-types appear. This section introduces these sub-types from the broader perspective of psychological testing; the following section traces their development in the specific field of language testing.

The most common sub-categories of validity are concurrent, predictive, content, construct and face validity. Reliability is sometimes seen as a necessary but not

sufficient for validity (Kline 1993; Underhill 1987). Many other types of validity have been proposed by different authorities, but in some cases these are merely different labels.

Concurrent and predictive validity involve comparisons against external yardsticks. Concurrent validity is claimed where a test performs well against another suitable measure contemporaneously, and predictive validity where it successfully anticipates future behaviour. Typically, such claims are based on statistical operations such as correlation, factor analysis and analysis of variance. A standard text on psychological testing considers that concurrent validity "is only useful where good criterion tests exist. Where they do not concurrent validity studies are best regarded as aspects of construct validity" (Kline, 1993:19) and predictive validity faces similar problems for lack of clear criterion measures, with the intervening time period as a source of additional variance.

Content validity examines the actual content of the test and compares it against the target behaviour that the test claims to measure overall. This is normally an intuitive process, although in some domains statistical issues of sampling and representativeness are raised. Content validity, Kline maintains, "is applicable only to a small range of tests where the domain of items is particularly clear cut. Tests of attainment and ability are of this kind". The example he gives is of a test of musical ability, which is a suitable candidate for content validity "because there is a good measure of agreement concerning the basic skills and knowledge, as is the case in language and mathematical ability". He concludes "In fact content validity is the validity to aim for, where it is

relevant, and it should be backed up with evidence of predictive or concurrent validity" (Kline, 1993:21-22).

Face validity refers simply to how it appears to its users; candidates and administrators, and more broadly the wider range of stakeholders such as sponsors and end-users of the information generated. Although derided for its lack of an empirical basis (e.g. Stevenson, 1981), face validity is often seen as a specialised validity criterion for performance assessment such as speaking tests, and may be known under other names, such as the terms transparency, meaningfulness and credibility mentioned in Messick (1994) and discussed in section 2.13 above.

Kline considers that face validity "is not related to true validity" because "subjects can guess what a face-valid test is measuring. Hence it is likely to induce faking or deliberate distortion..." This may be a problem for psychological tests in general, but for language tests, there is no need to disguise what the test is testing: quite the opposite, in fact, for motivational reasons. Kline acknowledges this by saying "In the case of ability tests, this is unimportant..."

It is construct validity that is ultimately acknowledged as the overriding form: "...construct validity embraces validity of every type". (Kline, 1993: 16-24) Construct validity asks whether the test matches the theory behind it. Problems occur in particular where this theory has not been made fully explicit. Where testing is used for research purposes, the interplay between construct validation and underlying theory is iterative; test data may lead to revision of the underlying theory, for example of language learning, and so to modifications of the constructs against which the test is to be

compared. This is particularly the case where statistical procedures are used for construct validation.

The American Psychological Association Standards

The American Psychological Association (APA) first published recommendations for educational and psychological tests in 1954, and has revised them as 'Standards' every ten years since (APA 1954, 1966, 1974, 1985 and most recently 1999). They provide a longitudinal source for Angoff's historical review 'Validity: an Evolving Concept' (Angoff, 1988), which describes the fundamental importance given to validity in psychometrics from the early days of the field, but considers definitions in the 1940s and 1950s to be in purely operational terms. "The view was held during the 1940s and even earlier... that the behavior of chief interest was the criterion behavior, and that it was left to the test author to develop a test, or to the user to find a test, that would predict that behavior" (Angoff, 1988: 21).

Predictive power, in particular, dominated the validation of psychology tests, with extensive use of multiple correlation to determine the validity of batteries of pen-and-paper tests against future performance on criterion tasks. Comparison against criterion behavior measured at the same time as the test became known as concurrent validity.

What exactly the tests measured was less important:

Whether or not the test performance measured psychological or educational constructs of interest as we define them today was of less importance than the fact that they correlated across the span of time ...Consistent with the empirical orientation and the emphasis on predictive validity that prevailed at the time, there was a strong tendency to think of criteria in strictly behavioral terms... (Angoff, 1988)

Angoff quotes the 1954 recommendations of the APA as identifying four types of validity: content, predictive, concurrent and construct. He also mentions the appearance of face validity in the literature of the 1940s, but concludes: "generally speaking, the effort to make a test face valid was, and probably is today, regarded as a concession, albeit an important one, to gain acceptability rather than a serious psychometric effort". The later 1966 and 1974 Standards "combined concurrent with predictive validity and referred to the two as a class, called 'criterion-related validity', reducing the number of types to three. In essence, then, validity was represented, even well into the 1970s as a three-categorized concept and taken by publishers and users alike to mean that tests could be validated by any one or more of the three general procedures". He also notes that by 1985, the standards referred to 'educational and psychological testing' rather than tests, a small but significant change reflecting a shift in the emphasis for responsibility for validation from the author or publisher of the test as a product to the use of the test as an activity.

Angoff describes the introduction of construct validity in the 1954 recommendations, under the influence of Cronbach and Meehl, as "a major innovation in the conception of validity and already perceived as the most fundamental and embracing of all the types of validity". Rather than a single validation activity, they proposed a "continuing research interplay to take place between the score earned on the test and the theory underlying the construct". All data, including empirical as well as personal and group data and the results of content analyses, are useful for construct validity:

"From the forgoing, we can see that construct validity as conceived by Cronbach and Meehl cannot be expressed in a single coefficient. Construct validation is a process, not a procedure; and it requires many lines of evidence, not all of them qualitative" (Angoff, 1988: 21-26).

A particular procedure for establishing an empirical basis for construct validation, Multitrait-multimethod (MTMM) first proposed by Campbell and Fiske (1959), is described in more detail in section 3.2.3 below.

The latest version of the APA Standards (1999) takes the evolution of validity a step further, by stressing that it is not the test itself that can be validated but the interpretation of test scores:

Validity is a unitary concept. It is the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed purpose. (APA, 1999: 11)

The burden of the accumulation of evidence for validation is to be shared between test developer and test user, with the user providing evidence derived from the particular context of use. Any new use of the test for any new purpose is a fresh interpretation and requires fresh validation. The notion of distinct sub-types of validity is dropped in favour of diverse sources of validity evidence. These include evidence based on test content; response processes; internal structure; relations to other variables; and the consequences of testing.

3.2 Validity in language testing

This section charts the evolution of the concept of test validity in the field of language testing. The categorisation and measurement of different sub-types of validity is discussed, and three broad strands emerge. These are:

- a) the different approaches to communicative testing that were simplistically characterised in chapter two as 'British/European' and 'North American' lead to a

tension between rationalist and empiricist approaches to test validation; the greater priority has usually been given to *post hoc* empirical validation by the north American school of thought, and to *a priori* rationalist validation by the British/ European school

- b) an increasing awareness of the importance of context, and a transition from validity as a context-free attribute of a test in any situation of use to validation as an activity that can only be meaningful when a range of contextual factors are taken into account. Most recently, new linguistic sub-disciplines of interaction analysis and conversation analysis have been applied to examine what is actually going on in a live oral test.
- c) in part as a corollary of b), validation has moved towards a continuing maintenance process consisting of a complex interplay of diverse sources of data rather than a one-time single procedure that establishes validity once and for all (and for all contexts). An analogy might be drawn with an ongoing quality assurance system that must be carried forward, in different manifestations, as long as the activity of which it is an integral part continues.

3.2.1 Validity as a single entity

Early text books on language testing tend to present validity as a single criterion. Lado (1961) does not explicitly distinguish between different types of validity, but refers to the linguistic content of the test, the situation or technique used to test this content, the contribution of non-linguistic factors, and correlation against other tests, as points to be considered. Valette (1967, 1977) similarly treats validity briefly, as one of the two

essential characteristics of a good test along with reliability. She mentions the duty on teachers to check both the content validity of a test and its validity against course objectives before use. She concludes:

For the language teacher, the degree of test validity is not derived from a statistical analysis of test performance, but from a meticulous analysis of the content of each item and of the test as a whole. (Valette, 1977: 46)

3.2.2 Validation of general language proficiency

One of the questions raised by the single global language proficiency hypothesis was what similarities this construct has with what other educational tests purported to measure, such as aptitude, scholastic achievement and personality inventories, and with that other general psychometric construct, intelligence. A corollary of an approach to test validation based primarily on empirical data is the ability to infer from such statistics the degree of overlap among superficially quite distinct measures. As many of the statistical tools were themselves derived from research into IQ, it is not surprising that language tests should be treated in the same way, as superficial evidence of deeper mental constructs whose interaction could be uncovered by the use of sophisticated statistical analysis. The most widely used statistical techniques were correlations and factor analysis, in various forms, to determine overlap with external criterion measures and patterns of internal variance between sub-tests. However, factor analysis and traditional item statistics such as item facility and item discrimination require complete data sets, and are unsuitable for analysing data from adaptive tests.

Using correlation data between language and IQ Oller concluded that many educational tests may in fact be measuring language:

the accumulating research evidence seems to suggest that a vast array of educational tests that go by many different names may be measures of language proficiency more than anything else. (Oller and Perkins, 1978: 14).

A content analysis of educational test questions (Gunnarsson, 1978) and a factor analysis of cloze test and educational achievement tests scores (Streiff, 1978) showed these tests to be measuring substantially the same thing.

From a critical point of view, these results can be considered trivial. Given that language is the medium through which almost all tests tasks are delivered and responses communicated, it would be surprising if this overlap didn't show up. Even so-called non-verbal tests require comprehension of instructions. By endorsing the existence of underlying general constructs, these results gave support to the integrative rather than the discrete-point approach to language testing, but also started to undermine the view of language as a closed system that could be independently measured. Reflecting the reality of language use in a wider context, Oller's important book *Language tests at school* (Oller, 1979) proposed a special sub-category of integrative tests which he called 'pragmatic'. A pragmatic test was defined as

any procedure or task that causes the learner to process sequences of elements in a language that conform to the normal contextual constraints of that language, and which requires the learner to relate sequences of linguistic elements via pragmatic mappings to extralinguistic context (Oller, 1979: 38).

This reference to a discourse level of analysis (language sequences sampled above as well as at or below the sentence level) and to the external, extra-linguistic context anticipated two of the key features of communicative testing. Indeed, the term

'pragmatic competence' was used as a label for one of the major subcategories of the over-arching language competence in Bachman's influential componential model (Bachman, 1990).

Oller categorises as 'pragmatic' apparently unrelated language tests such as dictation, cloze and composition according to the definition above, as an exercise in construct validation. He goes on to describe content and concurrent validity as the other major types, with the suggestion that reliability can itself be seen as a special case of the latter:

A special set of questions about concurrent validity relates to the matter of test reliability. In the general sense, concurrent validity is about whether or not tests that purport to do the same thing actually do accomplish the same thing ... Reliability of tests can be taken as a special case of concurrent validity. (Oller, 1979: 51)

Face validity only gets a footnote: "Such opinions are ultimately important only to the extent that they affect performance on the test. Where judgments of face validity can be shown to be ill-informed, they should not serve as a basis for the evaluation of testing procedures at all" (Oller, 1979: 52)

3.2.3 Construct validity as an empirical construct

In the North American approach to test validity, construct validation has traditionally been seen as a statistical activity, which can necessarily only be carried out after a test has been constructed at least in pilot form and sufficient data has been gathered for analysis and comparison with other appropriate measures of behaviour. It is a uniquely *post hoc* approach (Weir, 1990: 23)

An important collection of papers published in 1981 focused on 'The Construct Validation of Tests of Communicative Competence' (Palmer, Groot and Trosper, 1981) which brought together a number of attempts to apply current thinking on language test validation to the new construct of communicative competence, described in chapter two. Some of the contributions dealing specifically with oral testing are considered in the following section. In the introduction, Palmer and Groot identify three principal ways of evaluating validity: content validation, criterion-referenced validation and construct validation.

To investigate construct validity, one develops or adopts a theory which one uses as a provisional explanation of test scores until, during the procedure, the theory is either supported or falsified by the results of testing the hypotheses derived from it. This sequence ...will often be cyclical... (Palmer and Groot, 1981: 4)

Two specific empirical procedures are proposed for construct validation, both based on statistical correlations between tests. The first, described as 'a general procedure', involves five steps:

formulate a definition of the components of the communicative competence construct
locate existing tests or develop new ones to operationalise the provisional definition
form hypotheses and make predictions about the magnitude of the correlations between subjects' scores on the different tests
administer the tests
compare the obtained results with those predicted

In this procedure, some tests are predicted to correlate more highly than others, and "failure of the obtained correlations to conform to the predicted pattern would lead to the development of a new model (theory), or of tests which might be better operationalizations of the construct as previously defined, or of both" (page 5)

This general procedure was formalised into a system of convergent and discriminant validation. It was seen as a step forward from simple correlation between tests results, which only showed (in theory, at least) what similarity there was between what tests tested, that is how far they converged.

Convergent and discriminant validation is a logical extension of these traditional procedures; but since it requires evidence of discriminant as well as convergent validation it is ideal for the more rigorous, and functionally more important, problem of establishing the construct validity of language skill tests (Clifford in Palmer, Groot and Trosper, 1981: 62)

The second specific procedure proposed for construct validation is 'multitrait-multimethod validation' (MTMM), and much of the work in this collection is based on this technique, first described in Campbell and Fiske (1959). This postulates that any test score is a combination of the trait under measurement (e.g. communicative competence) and the method by which it is measured (e.g. dictation), and reflects the testee's ability both to communicate in the language and to complete dictation exercises.

The aim of the procedure is to allow a focus on the trait component by filtering out the effect of the method component. To do this, at least four different tests must be administered, combining at least two traits tested by at least two methods. In a similar way to the first procedure above, predictions are made and tested about the magnitude of the correlations between the results, but with two distinct kinds of validity being sought. 'Convergent validity' requires a high correlation between two tests supposedly measuring the same trait by different methods, while 'discriminant validity' seeks low correlations between tests of different traits measured by the same or different methods. Convergent validity is thus essentially the same as the usual procedure for establishing criterion-related validity, by comparing with other tests purportedly measuring the same thing.

Both these procedures suffer from their exclusive reliance on statistical correlations and assume that tests are measuring distinct skills. To the extent that this is not the case, what Carroll (1973) called the 'persistent problem' of an overall general language proficiency causes higher correlations between tests of different skills than might be expected, and undermines both the separate skills construct and the resulting data. Considerable statistical manipulation was needed to control for method-specific variance and to distinguish between method and trait (Bachman, 1990). Clifford concluded that

to be successful, a language skill validation study must use reliable assessment procedures and must be based on a language testing model which identified those aspects of language proficiency which overlap language skill areas (Clifford in Palmer, Groot and Trosper, 1981: 68-9)

There is again a risk of circularity here. Studies showing a substantial overlap between tests are fed back into theoretical model which is then used to provide construct validation for further statistical studies which support the model. This is a particular problem with correlation studies where there are so many possible interpretations of a single correlation coefficient between two tests scores.

An investigation by Bachman and Palmer (1980) specifically to explore construct validity by the multitrait-multimethod convergent-divergent design produced mixed results. They chose speaking and listening as their traits, as two maximally distinct aspects of language competence, and as their methods, interview, translation and self-rating. They found some evidence of discriminant validity, but they concluded that the design was too limiting to allow them to quantify the effect of method on test scores, and they recommended the use of "more powerful and enlightening" ways to research this, such as factor analysis. Discussing the use of MTMM for research into oral tests,

Fulcher highlighted the risk of cross-contamination between scores on two or more traits which are operationalised through rating scales, with the same raters giving a score on one scale and then carrying that score over to another scale (Fulcher, 1994: 39).

From a rationalist point of view, requiring different traits to be tested by different methods might be thought to be a recipe for confusion; being 'maximally different', speaking and reading would obviously be best tested by different methods.

Stevenson quotes Cronbach and Meehl as suggesting that construct validity "must be investigated wherever no criterion or universe of content is accepted as entirely adequate to define the quality to be measured" (Cronbach and Meehl, 1972:92) but acknowledges that there is no single well-established procedure for establishing it, compared with the routine use of correlation used for criterion-related validity. One of the difficulties in applying MTMM to language testing is that "what is trait and what is method is very hard to distinguish, and what should be considered as trait and what should be considered as method is very hard to decide" (page 53). Stevenson gives the example of an oral interview, where successful participation could be viewed as an ability either to communicate (trait) or to participate in oral conversations (method). He suggests this is a central problem in language testing, in particular, because "we are trying to measure something with tools that are made largely out of what we are trying to measure, and the problem is to separate the tool from the matter". (page 54)

Although the empiricism of the purely statistical approach to construct validation has an attractive objectivity, it does not help the test developer to design a new test based on

the theoretical principles of a particular approach. Even where a detailed theoretical framework may be missing, such as is the case with the communicative approach, this 'does not absolve test developers from trying to establish *a priori* construct validity' (Weir, 1990: 23) which can subsequently be tested out by statistical comparisons.

3.2.4 Validation in the rationalist tradition

In the more pragmatic testing tradition in the UK, communicative language teaching also revolutionised language testing, but with different consequences. Crudely speaking, the north American school characterised by Palmer et al. (1981) and Stevenson (1981, 1985a, 1985b) above carried on the unquestioned assumption that a *post hoc* psychometric approach is the only scientifically respectable basis for test validation. The British school, represented by Davies (1968), Heaton (1975, 1988) and Morrow (1977, 1979), saw a tension between the need to measure language in use performance and the requirements of the statistical approach to validation, and in particular to reliability indices as one form of validity evidence. Davies (1978) called it the reliability-validity tension; Underhill (1982) as a trade-off, where you could design a test that represented a compromise at any point along a continuum between high reliability and low validity at one extreme, and high validity and low reliability at the other.

Because of the increasingly sophisticated statistical techniques, testing in the empirical tradition became increasingly the preserve of the expert; writing specifically for language teachers about construct validation by correlation, Hughes wrote

Construct validation is a research activity, the means by which theories are put to the test and are confirmed, modified or abandoned. It is through construct validation that language testing can be put on a sounder, more scientific footing. ... but it will not happen overnight.... When in doubt, where it is possible, direct testing of abilities is recommended. (Hughes, 1989: 27)

Of five types of validity - face, content, predictive, concurrent, and construct - Morrow felt there was an internal circularity underlying at least three:

... with two exceptions (face, and possibly predictive), the types of validity outlined above are all circular. Starting from a certain set of assumptions about the nature of language and language learning will lead to language tests which are perfectly valid in terms of these assumptions, but whose value must inevitably be called into question if the basic assumptions themselves are challenged. (Morrow, 1979: 147)

This criticism is more broadly and obviously true of any theoretically based validation exercise; a test may not measure up against a criterion which is outside the construct on which the test is based. It is a fair critique of the theory underlying the discrete point approach in a communicative context, but not of the construct validity of test based on it. Morrow concludes

There is clearly no such thing in testing as 'absolute' validity. Validity exists only in terms of specified criteria, and if the criteria turn out to be the wrong ones, then validity claimed in terms of them turns to be spurious. (Morrow, 1979: 147)

In fact, all that Morrow is doing is pointing out that with the advent of the communicative approach, the theoretical horizon had changed considerably, and therefore that previous criteria could no longer be sustained.

Morrow's specific conclusions with regard to validity were that communicative testing would be characterised by

modes of assessment which are not directly quantitative, but which are instead qualitative. It may be possible or necessary to convert these into numerical scores, but the process is an indirect one and recognised as such... Reliability, while important, will be subordinate to face validity. Spurious objectivity will no longer be a prime consideration, although it is recognised that in certain situations test formats which can be assessed mechanically will be advantageous. (Morrow, 1979: 150-151)

This subordination of reliability to face validity is diametrically opposed to Stevenson's view, above.

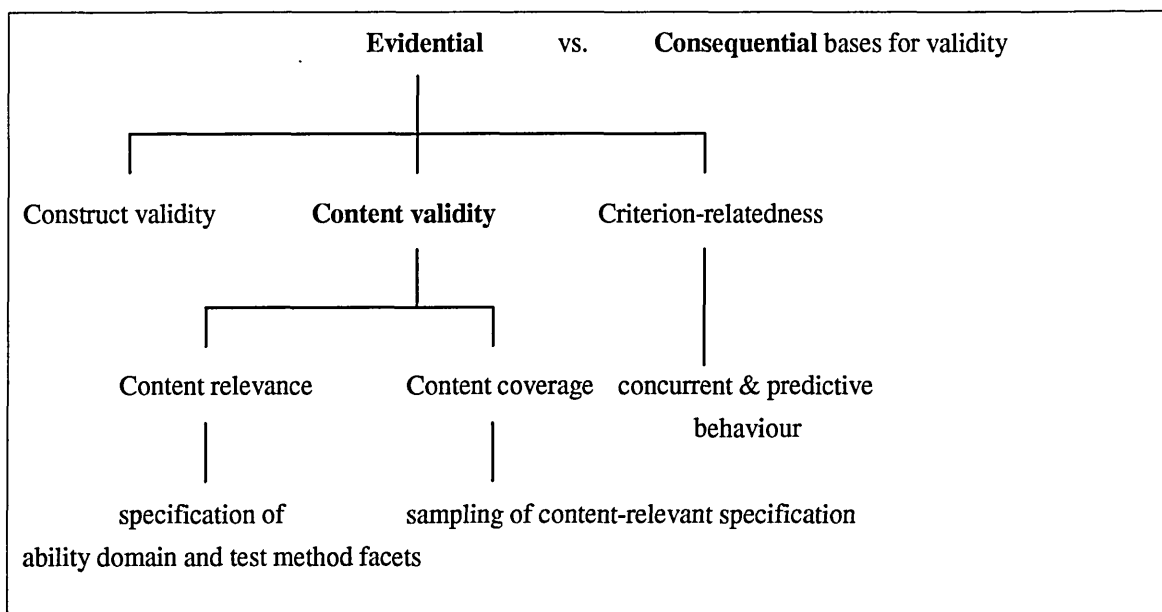
Quantitative assessment here refers to a more or less mechanical process of counting correct answers. Qualitative assessment by contrast refers to an active marking process, where the assessor compares the performance observed against certain pre-defined criteria such as speed, range, accuracy, repetition of the candidate's performance, which may be scored individually and then summed to give an overall score. Or alternatively they may be combined in a series of profiles or band descriptors from which the marker chooses on an impressionistic basis. In practice, both these approaches have been adopted by different tests. A major problem for both marking systems is doubt about the consistency of judgements of assessors; and the validity as well as the reliability of the early band descriptors has more recently been criticised for their armchair origins and lack of theoretical underpinnings (e.g. Fulcher, 1987) and North and Schneider have recently demonstrated how IRT can be used to validate proficiency scales (North and Schneider, 1998).

3.2.5 Content validity

Although content validity is universally acknowledged as a core sub-type of validity, few authors give specific procedures for establishing it.

Bachman (1990) divides content validity into two aspects: content relevance and content coverage. Content relevance in turn comprises the specification of the ability domain and of the test method facets. The ability domain is the behaviour sought from the participants, the specification of which Bachman describes as "essentially the process of operationally defining constructs". Thus content and construct validity overlap substantially in providing the evidential basis for validity. The test method facets describe the context : "Every aspect of the setting in which the test is given and every detail of the procedure may have an influence on performance and hence on what is measured" (quoting Cronbach, 1971: 449) including, for example, the physical conditions of the room, the seating arrangement, personality characteristics and so on. Content coverage is a question of appropriate sampling of the tasks required once the behavioural domain has been adequately described. Bachman's (1990) model of validity is illustrated schematically in Figure 3.

Figure 3 **Schematic diagram of Bachman's (1990) model of validity**



In practice, Bachman acknowledges that the complete specification of the behaviour domain is impossible, except in very restricted cases, and without proper specification the question of how to sample cannot be answered satisfactorily either. A second major weakness of content validity is that it takes no account of how individuals perform (Bachman, 1990: 247; Brown, 1996: 239). It focuses on tests, not on test scores or how they are used, and can therefore only ever provide partial evidence for general validity.

Alderson et al. (1995) identify content as one of three main types of validity in the language testing literature: rational, empirical and construct.

Rational or “content” validation depends on a logical analysis of the test's content to see whether the test contains a representative sample of the relevant language skills. Empirical validation depends on empirical and statistical evidence as to whether students' marks on the tests are similar to their marks on other appropriate measures of their ability... Construct validation refers to what the test scores actually mean. (Alderson et al 1995:171)

However, they suggest that the rational /empirical distinction is no longer useful, and propose instead the terms 'internal' and 'external' validity, where external validity is what the American Psychological Association Standards (APA, 1985) refer to as 'criterion validity'. Both internal and external validity may draw on empirical data. 'Internal validity' has as sub-types face validity, content validity and response validity, the latter involving, for example, introspective accounts of the test-taking. 'External validity' consists of concurrent and predictive validity, and most frequently draws on correlation studies.

Expert panels and moderating teams

One technique for establishing content validity is the use of an expert panel or moderating team. There are few detailed descriptions of the use of 'expert panels' in the testing literature; two specific examples are described in section 5.1.2.

Brown (1996) suggests convening an expert panel as one way of exploring content validity. He advises that

unfortunately, this procedure is only accurate to the extent that the bias of the experts do not interfere with their judgments. Hence, test developers may wish to take certain steps to ensure that the experts' judgments are as unclouded as possible. (Brown, 1996: 234)

One of these steps is to ensure that "at least to a degree, the experts share the kinds of professional viewpoints that the testers and their colleagues have." (page 235) The example he quotes, of a shared empathy for the communicative approach, suggests that he is thinking of agreement on quite fundamental issues, at the level of underlying construct rather than content. The obvious risk is that experts are selected precisely because it can be anticipated that they will not disagree.

Alderson et al. (1995) suggest that a moderation committee may meet the requirements for content validation, but only if they have genuine expertise in the field and their judgments are collected in a systematic way, through a data collection instrument, in order to gain some idea of the degree of consensus. In their experience, this rarely happens, and instead committee members pool opinions informally and unsystematically, and as a result the group dynamics may well affect the outcome (page 174). It was to avoid this kind of peer pressure that the Delphi technique for collecting anonymous consensus was used in the present study, and this is described further in

chapter five. Both Alderson et al. (1995) and Brown recommend the use of scales to rate test items against, rather than seeking yes/no judgments of validity.

Crocker and Algina also recommend a panel of independent experts, and say explicitly that they should not be the same people involved in writing the test items (Crocker and Algina, 1986: 220). They discuss various quantitative indices for summarising judges' decisions in a context where the sought-after performance domain is specified by a list of instructional objectives, but they do not address the issue of consensus and peer influence or the bandwagon effect at all.

Some sources recommend the use of expert advice, but without constraining it to the panel format. The latest American Psychological Association Standards recommend "expert judgements of the relationship between parts of the test and the construct" as a source of evidence based on test content (APA, 1999: 11) and Weir (1990) suggests inviting professionals in the field to comment on the texts, formats and items in a pilot test as "a further validation check" (Weir, 1990: 39)

For checking individual items rather than content validation of the test as a whole, many sources suggest that internal moderating of proposed items among a team of developers is sufficient without external expertise, but Hughes warns of the dangers:

Colleagues must really try to find fault; and despite the seemingly inevitable emotional attachment that item writers develop to items that they have created, they must be open to, and ready to accept, the criticisms that are offered to them. Good personal relations are a desirable quality in any test writing team. (Hughes, 1989: 51)

Differentiation of the expertise offered by panel members may be desirable. Crocker and Algina recommend asking qualified colleagues to review items and suggest that this

item review panel should consciously be selected to include different skills: as well as subject matter experts, there should be expertise in measurement and test construction and one or more members should have 'expert familiarity with the population for whom the test is intended' (Crocker and Algina, 1986: 882)

3.2.6 Validity as a unified concept

Most recently, construct validity has emerged as the overriding form of validity, with other types feeding into it. Cumming lists sixteen different types of validity, and concludes,

Rather than enumerating various types of validity as appear above, the concept of *construct validity* has been widely agreed upon as *the* single, fundamental principle that subsumes various other aspects of validation (i.e. those listed above), relegating their status to research strategies or categories of empirical evidence by which construct validity might be assessed or asserted (Cumming in Cumming and Berwick, 1996: 5)

He goes on to quote similar statements from the Standards from the American Psychological Association (APA 1985), defining construct validity as "the most important consideration in test evaluation", and Moss (1992: 232) defining validity as "a unitary concept requiring multiple types of evidence to support specific inferences made from test scores".

One of these types of evidence is from what has traditionally been called reliability, but the apparently clear-cut demarcation is not always as easy as it seems, particularly where statistical tests involved, such as correlation or factor analysis, might be interpreted as data for either reliability or validity. Bachman (1990) suggests we view the distinction more as a continuum, and accepts that test content and test method are

inextricably bound together, making it ultimately impossible to distinguish reliability and validity.

Another type of evidence that is brought into the new unified concept of validity is the broader test context, including the interpretation and use made of test scores, and subsequent actions that follow. This inclusion of consequences introduces an ethical dimension into validation.

This signals a crucial move from validity as an attribute of a particular test, which may then be transferred to another quite different situation, to validity as an attribute of the test use in a particular context. It is a move from validity as a trialling activity prior to the launch of a new product to validity as something that can only be established in regular use; and it is therefore a move from validation as a one-time activity to validation as a continuing process.

"Validity is an overall evaluative judgment, founded on empirical evidence and theoretical rationales, of the adequacy and appropriateness of inferences and actions based on test scores" (Messick, 1988: 33)

A new distinction can therefore be made between evidential bases for test validity, such as empirical appraisals of the relationships between constructs, and consequential bases for test interpretation, such as judgmental appraisals of the outcome of a test: "a value context of implied relationship to good and bad, to desirable and undesirable attributes and behaviors" (Messick, 1988: 41). Validity extends to the use made of a test, and test designers therefore have an ethical responsibility to consider how test scores are likely to be interpreted, and test users to validate the interpretation of individual scores from individual test events.

The evidential/consequential distinction was illustrated in Figure 3. Components of the evidential basis are content, criterion and construct. Content includes content coverage, that is adequacy of sampling of the target behavioral domain, as well as content relevance. Criterion relatedness includes data from concurrent and predictive behaviour. Evidence supporting construct validity may be from correlational evidence, including factor analysis and MTMM, and from experimental evidence, by pre- and post-testing control and experimental groups, but both of these sources take as their input data only the product of the test, that is, the test scores. A third source of evidence is an analysis of the test taking process itself, perhaps including self-reporting data, exploring the strategies used by test-takers and asking to what extent and why different individuals perform a task in different ways.

Consideration of the consequential basis of validity involves moving 'out of the comfortable confines of applied linguistic and psychometric theory and into the arena of public policy' (Bachman, 1990: 281). It is a much more difficult area to be precise about; questions include balancing the conflicting demands of reliability, validity and practicality; potential differences in the value systems of tests designers, administrators, test takers and other stakeholders; robustness of the test for so-called 'high stakes' testing, where it may be used to influence the future academic or professional careers of test takers, with little or no recourse or right of appeal.

A list of the possible consequences of test use, both positive and negative, is likely to feature 'washback' or 'backwash', the influence of test content and design on the teaching and learning styles of the instructional programmes of which the tests are a

part. Reference to the value systems of test takers, particularly where these may differ from the cultural background of the designers, raises again the question of face validity.

The inclusion of the consequences of testing in the validity area has been controversial and does not enjoy unanimous support. Sources cited by Busch argue that it clutters and confuses the concept of test validity, and that in practice it is driven by a response to substantially increased test-related litigation in the United States (Busch, 1997). Among information published about the computer-based version of the TOEFL test is the statement that 'because TOEFL is committed to the highest standards of test design, fairness, reliability, validity and security, institutions of higher education can be confident that TOEFL scores provide measurement information that is accurate and legally defensible' (ETS 2000)

Most recently, 'usefulness' has been proposed by Bachman and Palmer as an overarching function of several different test qualities, including validity and reliability:

$$\text{Usefulness} = \text{Reliability} + \text{Construct validity} + \text{Authenticity} + \text{Interactiveness} + \text{Impact} + \text{Practicality} \text{ (Bachman and Palmer, 1996: 18)}$$

The term 'impact' here includes consequential validity. The implication is that 'usefulness' is a superordinate of validity; however much validity is seen as a unified construct, with multiple sources of data feeding into it, there is a higher-order construct above and behind it. The individual components cannot be evaluated independently, but must be seen as contributing to the whole, and it is the overarching 'test usefulness' that is to be maximised, rather than individual qualities. In the design of a model for continuing test validation, this concept of usefulness is picked up in chapters 7 and 8 as a convenient label for programme evaluation that is broader in scope than validation.

Table 2 summarises the of the major models of validity described in this section, and distinguishes the rationalist from the empirical components of each.

Table 2

Summary of major models of validity

Source	Rationalist components	Empirical components
Lado (1961), Valette (1967)	Validity is largely undifferentiated	
APA (1966, 1974), summarised by Angoff, below	Content, construct	Criterion (concurrent + predictive)
Angoff (1988), quoting APA (1966, 1974)	"In essence, then, validity was represented, even well into the 1970s as a three-categorised concept and taken ... to mean that tests could be validated by any one or more of the three general procedures"	
Morrow (1979), Hughes (1989)	Face, content, construct	Concurrent, predictive (criterion)
Oller (1979)	Construct, Content	Concurrent; reliability is special case of concurrent validity
Palmer, Groot and Trosper (1981)	Content	Construct, via multitrait-multimethod and convergent-discriminant methods
Underhill (1987)	Face, content, construct	Concurrent, predictive, and reliability as a necessary but not sufficient condition for validity
Messick (1988)	Evidential vs consequential bases for validity	
Moss (1992)	Validity is 'a unitary concept requiring multiple types of evidence to support specific inferences made from test scores'.	
Alderson et al. (1995)	The rationalist /empirical distinction is no longer useful, and they propose instead 'internal' and 'external' validity, where external validity is what the APA Standards (1985) refer to as 'criterion validity'. Both internal and external validity may draw on empirical data. 'Internal validity' has as sub-types face validity, content validity and response validity, the latter including e.g. introspective accounts of the test-taking. 'External validity' consists of concurrent and predictive validity, and most frequently draws on correlation studies.	
Cumming, in Cumming and Berwick (1996)	'...the concept of <i>construct validity</i> has been widely agreed upon as <i>the</i> single, fundamental principle that subsumes various other aspects of validation (i.e. those listed above), relegating their status to research strategies or categories of empirical evidence by which construct validity might be assessed or asserted.'	
Bachman and Palmer (1996)	Usefulness = Reliability + Construct validity + Authenticity + Interactiveness + Impact + Practicality	
APA Standards (1999)	'Validity is a unitary concept. It is the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed purpose.'	

3.3 Validity issues in oral testing

This section considers some validity issues that have special relevance for the assessment of speaking, reflecting the distinguishing features of communicative language testing described in chapter two above. They are taken up again in chapter seven where their implications for adaptive, computer-based tests of communicative skills are discussed.

3.3.1 Direct and authentic

The validation of oral tests is inextricably bound up with the terminology of 'communicative', 'direct', 'interactive' and 'authentic'. They are sometimes used interchangeably by some authors (Bachman, 1990: 301), and overlap to a greater or lesser extent for others, and it is not possible to demarcate their meanings conclusively. 'Communicative' is defined, more or less, in chapter two; there is no single criterion, but enshrine a small number of widely-accepted principles. 'Authentic' is a direct synonym for 'lifelike' or 'real world', and is used primarily in the context of the communicative approach to describe teaching materials; thus, a leaflet produced for commercial or publicity purposes in the real world, and introduced into the classroom to be exploited for its language, would be deemed 'authentic'. A radio bulletin which is based on a real excerpt, but is re-recorded and perhaps slightly re-written, might be called 'semi-authentic'. A test can be direct, by measuring precisely what we want to measure, but is unlikely to be really authentic, because candidates know it is a test and not real life (Hughes, 1989: 15).

'Direct' was first used as a label for tests by Clark (1975):

In direct proficiency testing, the testing format and procedure attempts to duplicate as closely as possible the setting and operation of the real-life situations in which the proficiency is normally demonstrated (Clark, 1975: 10)

whereas there is no requirement on indirect tests to reflect situations of authentic language use. The original analogy is that of a test going directly to the heart of what is to be measured, rather than approaching it indirectly; an indirect test of speaking might claim to measure speaking ability by virtue of concurrent correlation with a direct test:

The fact that correlations proved to be consistently high between LPI [Language proficiency interview, a genuine oral test] and TOEIC-LC [a multiple-choice test of listening] strongly suggests that both tests are, in fact, effectively measuring the common ability to understand and use spoken English. (ETS, 1993: 11)

The underlying assumption is that the target language competence is evidenced in real-life language use, and that 'directness' or authenticity is a contributory factor to test validity. For some authors, the relationship is more direct: the construct of language proficiency is operationalised as language in use in authentic contexts, and therefore the content, construct and to an extent face validity of an oral test can be interpreted in terms of the extent to which it mirrors real life language use (Clark, 1975, 1978; Lee et al., 1985).

This strong claim that direct tests are inherently valid to the extent that they reflect real life is disputed by, for example, Messick (1981) and Cronbach (1988) and described by the former as 'operationism' (see also the quote from Cronbach in Messick, 1994: 20). This is a confusion of a sample of behaviour (the actual language sample elicited) with the ability itself, and fails to distinguish between test and construct.

Messick's theoretical starting point for validity is the centrality of construct validity in all forms of assessment (Messick 1981, 1988, 1989) and in an influential article in 1994 he focuses on the use of the terms *authentic* and *direct*: "the portrayal of performance assessments as *authentic* and *direct* has all the earmarks of a validity claim but with little or no evidential grounding". He examines the way the terms are used and concludes that they can be accommodated into the theoretical framework of construct validity, as the overriding central form of validity, and viewed in terms of construct representation.

Construct representation is the extent to which the test tasks successfully map onto the target domains, with two possible areas of mismatch. Where there are areas of the target domain which the test does not tap, there is a problem of construct under-representation, and where the test assesses other factors which are not in the target domain, then there is a problem of construct irrelevance. From this perspective, a concern for authenticity can be seen as an attempt to minimise construct under-representation through a desire to 'leave nothing out'. It is a response to the concern that decomposing a complex skill into components for separate measurement will fail to capture an important part of the criterion behaviour, which will then be undervalued, untaught and so underdeveloped.

Messick concludes that "this perceived devaluing of complex skills in favor of component skills in educational testing, and ultimately in teaching, is what energizes the [authentic testing] movement" (1994: 20). This is ultimately driven by concern about the backwash effect of testing, and is perhaps symptomatic of a lifetime of research in educational and psychological testing in the American educational system; he is writing

about performance tests in education in general, but his views have been highly influential in language teaching.

Directness, on the other hand, is interpreted in this framework as an attempt to minimise construct irrelevance in the process of assessment, specifically method variance induced in test scores from such sources as candidates' testwiseness in coping with different item types and differential guessing. Instead of a concern about 'not leaving anything out', this is seen as a concern about 'not putting anything extra in', but is interpreted uniquely in terms of directness of assessment, with the implication that direct assessment is more likely to employ open-ended tasks and judgemental scoring (Messick, 1994: 21). This is a much narrower view of directness as a validity criterion.

'Authenticity' is one of the five test qualities that are included in the Bachman and Palmer's superordinate term 'usefulness', quoted in 3.2.6 above (Bachman and Palmer, 1996: 18)

For Spolsky, it is not so much the actual nature of the task that is necessarily inauthentic as the 'breach of a normal conversational maxim' that is fatally flawed because it asks the candidate to behave in an unnatural way: "to answer questions asked by someone who doesn't want to know the answers because he already knows them" (Spolsky, 1990: 10-11). Some of the techniques suggested by communicative testing and used in the Five Star test, such as information gap tasks and the personalisation of tasks to elicit an individual's experiences and views, get round this artificiality, but the fact remains that there is a 'social contract' between interviewer and candidate to accept the rules of the game. Spolsky had previously elaborated a contrast between 'authentic test language'

and 'real life language', in terms of the goal of the interaction to obtain an assessment, the participants' previous lack of acquaintance, the decontextualised setting, the choice of topic imposed by the interviewer, and the fixed time limit. (Spolsky, 1985; Stevenson, 1985a)

Localisation

A different perspective on authenticity is localisation, the extent to which a test is designed to reflect the particular context in which it is to be used, whether that context is cultural, social, professional, educational or any combination of these. This reflects the communicative criterion of contextualisation, linking meaningful test tasks to target language use, with the implication that no one test can match all situations of use (Weir, 1990, quoted in section 2.9.2). The related criteria of individualisation and topicality are considered further in chapter seven.

While the English tests and examinations from the major English language examining boards - ETS in USA and UCLES in UK for example (see glossary) - are designed and validated to be globally applicable, a common feature among smaller-scale language test developers has been the need to develop tests that are sensitive and appropriate to particular local needs and circumstances. The type of test for which a validation process is being developed here is for such a small or medium-scale test which has some specificity to the geographical or cultural target market.

One of the implications of Messick's consequential basis for validity is a heavy responsibility on test users for the interpretation of test scores in the social and political context in which it is actually used.

The test user is in the best position to evaluate the meaning of individual scores under local circumstances, that is, to appraise the construct interpretation of individual scores and the extent to which their intended meaning may have been eroded by contaminating influences. (Messick, 1988: 43)

This burden does not fall on the test user alone, and the test designer or publisher cannot avoid their share of the responsibility, and must anticipate the range of contaminating influences and possible values and consequences of test use. One resolution is a truly local development, where the test designer is already a part of the context for which the test is intended and in which it is to be used.

However, such local specificity will reduce external reliability and make comparisons with results from other contexts difficult to interpret. Global tests have an advantage in this respect of having a clearly-defined market - the whole world - and so being able to establish research designs to collect and compare data from different sectors and regions to build a picture of the patterns of variation across and within local markets.

The native speaker as model

Language testing has traditionally used the native speaker as the model against which to measure candidate performance. This yardstick may be explicit in the rating scales, for example "handles all oral interaction with confidence and competence similar to those in own mother tongue" (Carroll and West, 1989: 22) or implicit in the use of native speaker interviewer/interlocutors only, as is the case with the Trinity College Spoken English Grade Exams. The rationale underlying this is perhaps summed up by the 'foreign travel' metaphor; learners are people who come from their own culture to the foreign country and they need to be able to communicate with and survive among the natives.

Objections to this model are both theoretical and practical. The recently developing field of intercultural communication raises questions about the theoretical basis, arguing that it perpetuates the existing social hegemony and denies individual learners the right to use the language for themselves rather than imposing behavioural norms from another culture (Andrews and Fay, 1999: 391-2). Practical problems posed are the near impossibility of defining 'native speakerness' (Lantolf and Frawley, 1985) and evidence that native speakers do not themselves perform consistently well on language tests, or even consistently better than non-native speakers (Bachman, 1990: 248-249).

A large and increasing proportion of the use of English is as an international language, in other words, for communication between two or more non-native speakers for whom it is the only, or the best, common medium.

Constructs of proficiency founded on 'native-like' prescriptions fail to account for regional and international influences, and these can affect aspects of language proficiency from 'world-knowledge' to interlanguage features (Pollard and Underhill, 1996: 49).

While many aspects of the target language use may be difficult to define in practice, it is generally not difficult to determine the linguistic background of the other participants that candidates will interact with, and the communicative criterion of authenticity would require that a test reflect this.

Overall, the nature of authenticity remains problematic, and this is explored further in chapter seven.

3.3.2 Interaction

Oral testing is unique among types of language test in that it typically involves live communication between participants in 'real time', in other words, as the test takes place. This interaction was identified at the outset as one of the key features of communicative testing (section 2.9.2 above) and this 'face-to-face talk' is the first of van Lier's basic prerequisites for evaluating oral proficiency (van Lier, 1989) along with decision-making opportunities and goal-relatedness, discussed in section 2.9.4 above.

In some situations, the term 'interaction' is used as a label for a speaking test, to emphasize this two-way process. The Dutch Ministry of Education has deliberately chosen to call the speaking component of its school language tests 'oral interaction ability' rather than 'speaking ability' (Wijgh, 1993). The Cambridge Exams Syndicate's Certificates in Communicative Skills in English speaking component was formerly called 'oral interaction' (UCLES, 1995) but has recently been changed in part because the exam has a separate listening component.

A simple everyday interpretation of 'interaction' in the context of an oral test might be two-way spoken communication between two or more participants, such as interlocutor and candidate, and this is the criterial feature in distinguishing between direct and semi-direct tests (see glossary, and further discussion in 7.2.1). The interaction need not all be face-to-face; communication by telephone may be used for reasons of validity or practicality, and interaction can take place in writing as well as in speaking, with message-taking and giving tasks requiring the use of combined skills to reflect authentic target language use. A more narrow interpretation of 'interaction' (Weir, 1990: 78) is

restricted to tasks where there is an information gap between participants, for example, where the candidate has to elicit information from the interlocutor which s/he does not already know.

Although the prototypical oral interview takes place between a single interviewer and a single candidate, variations allow for more than one candidate, or more than one interlocutor/ assessor, and for a range of patterns of interaction to take place between these participants. For example, all but the lowest level of the five UCLES main suite exams (UCLES, current) will soon contain face-to-face paired speaking tests, in which two candidates engage in different patterns of interaction with an interlocutor in different tasks. A second examiner acts as assessor only and plays no part in the interaction.

The last of the Cambridge 'main suite' exams to be revised in this format is the highest level, the Certificate of Proficiency in English. Currently, all the interaction in the CPE oral interview lasting approximately 15 minutes takes place between a single candidate and a single interviewer who combines the roles of interlocutor and assessor. The new format of the revised speaking test to be introduced in 2001 consists of an interview between each candidate and the interlocutor, very similar to the current format; a collaborative task between the two candidates; and an individual 'long turn' presentation by each candidate followed by discussion between both candidates and the interlocutor.

The structure of the revised speaking test is based on recent research (Ffrench, 2000) that compared the range of language functions employed by candidates in the old and the new test formats. The research used an 'observational checklist' of 30 functions

grouped into three categories, informational functions, such as providing personal information, speculating, elaborating, describing and comparing; interactional functions, such as agreeing, disagreeing, persuading, asking for opinion or information; and functions for managing interaction, such as initiating, terminating or changing topic.

A straight count of how many functions were used altogether showed that three candidates in the old individual interview format each used 14 of the total of 30 possible different functions, while three candidates tested in the new paired format used 25, 26 and 27 of the possible 30 functions. An analysis of the proportional breakdown across the three function categories suggested that candidates in the paired test format used language much more for managing interaction and interactional functions than candidates in the individual interview format. Overall, the conclusion was that the paired speaking test offers candidates the opportunity to produce a much richer sample of language (Ffrench, 2000). What the research does not do is to claim that this richer sample is necessarily more lifelike, in the sense of more representative of real-life target language use; and as the Cambridge main suite exams are general purpose tests of proficiency, used globally for a very wide range of purposes, it would be extremely difficult to claim to define the target language use with any precision.

There are however several criticisms of paired speaking tests, which focus on the new sources of bias that the format introduces, compared to the single candidate/interviewer format. These additional sources of bias include the degree of the candidates' familiarity with each other, or lack of it; differences of age, personality, status and whether or not they share a common mother tongue; the extent to which their proficiency levels are

matched, and for this and other reasons, the evenness of their contributions to the interaction (Foot, 1999a).

There is also some evidence to suggest that in order to be mutually comprehensible in the interests of task achievement, participants in paired tests may actually adjust their pronunciation and language:

candidates with different first languages tended to engage in a kind of foreigner talk. Those sharing the same first language achieved mutual intelligibility by exaggerating their first language accent (Foot, 1999b, citing Jenkins, 1997)

Interaction between participants and feedback on relevance and correctness of response are the distinguishing features of what Bachman and Palmer call reciprocal tasks, "so that the language used by the participants at any given point in the communicative exchange affects subsequent language use" (1996: 55). This influence of earlier language use on later language use in the test is contrasted to an adaptive test, where the actual selection of later tasks, rather than the language used, is determined by responses to earlier ones. Most adaptive tests are therefore interactive, but because test takers do not receive feedback on the correctness of their responses, they are not reciprocal. A situation in which an interlocutor simplifies or paraphrases a task which has not been properly completed would be interactive and adaptive; it would also be reciprocal, if there is an element of feedback on performance. Ultimately, it could be difficult in live conversation to determine whether feedback had taken place, and different participants might have different interpretations as to whether feedback had been intended and/or perceived.

There is a contrast between this everyday use of 'interaction' and 'interactiveness', which is another of the five test qualities in Bachman and Palmer's superordinate term

'usefulness' quoted above. Their definition of this latter term rests on the interaction between candidate and task, rather than candidate and interlocutor or any other person:

We define *interactiveness* as the extent and type of involvement of the test taker's individual characteristics in accomplishing a test task. (Bachman and Palmer, 1996: 25).

It is features such as topical knowledge, affective schemata and metacognitive strategies that are called into play when a task is interactive, in this sense, and there is, as they point out, a vital link here with construct validity. There appears to be no requirement for participation by any other person.

Attention has focused in the last few years on the nature of the interaction that takes place in oral tests, and in particular on the extent to which it can be said to reflect authentic target language use. Different theoretical frameworks have been employed to examine test interaction in some detail, and the general consensus is that there are significant differences between the interaction in an oral test and in everyday conversation (Young and He, 1998; McCarthy, 1991).

Three of the theoretical frameworks used to analyse oral interaction are

1. speech acts (Austin, 1962; Searle 1969, 1979) and Gricean maxims (Grice, 1975, 1989).
2. speech events (Hymes, 1970) or speech activities (Gumperz, 1982), reflecting a concern from social anthropology stretching back to Malinowski (1923) to perceive the meaning of interaction not just at the level of individual utterance but in the broader culture-specific context of situation.

3. conversation analysis (the work of Schegloff and Sacks, e.g. Sacks, Schegloff and Jefferson, 1974) looking at such features as turn-taking, length of turns, and topic nomination in the interaction.

The use of the speech acts tradition has been criticised in cross-cultural literature for unjustifiably assuming a universality of speech functions across cultures, and so perpetuating a linguistic singularity that is anglocentrically-dominated (Wierzbicka, 1985; Clyne, 1996 cited in Andrews and Fay, 1999). The use of a common language should not be taken to imply that it must always be used in the same way.

A study in the conversation analysis tradition that examined an oral proficiency interview (OPI, a widely-used standard format used by different testing agencies in the United States) found a number of differences from ordinary conversation (Johnson and Tyler in Young and He, 1998). Compared to a normal conversation which has systematic turn-taking and roughly equal distribution of length of turns, the test showed an extreme imbalance in the length of turns with the candidate speaking far more than the interviewers, and the interviewers sometimes failing to take turns.

Instead of spontaneously created and negotiated topics, the analysis of the test revealed the pre-determined agenda of the interviewers, and their determination to elicit certain types of language from the candidate, such as description, which led to a loss of coherence in the interaction (this could also be interpreted as violating the Grice's Cooperative Principle). The great majority of information questions were posed by the interviewers, and the candidate tried to answer all of these, but of the few that she asked, the interviewers attempted to respond to only one. The researchers interpreted

this lack of responsiveness as a failure of conversational involvement on the part of the interviewers.

Their conclusion was that the OPI could not be considered a 'valid example of a typical, real-life conversation' and that it 'may very well be a unique speech event with its own unique norms and rules' (Johnson and Tyler in Young and He, 1998: 28-31).

This study was based on the analysis of a single oral test, using a test framework that sets down quite rigidly what type of language the interviewer should use as well as what type of task at different stages of the interview: "The prescribed format of the interview, not the emergent discourse, controls both the local and overall structure of the exchange" (page 44). In part this is because the marking system of this OPI test requires it; at one point, for example, the exchange becomes uncomfortably like an interrogation as the interviewer repeatedly attempts to elicit an answer to a Supported Opinion Question such as 'Why do you think so?' or 'What do you think are the reasons for that?' which would count as evidence of a Level Three rather than a Level Two performance.

Repetition of the analysis with other candidates and interviewers using the same test might not yield the same results, and an analysis of test events using other less rigidly prescribed test formats might reveal different features. The dominance of the question-and-answer interview technique as the single method might be expected to produce a very black-and-white picture, compared to oral tests that deliberately employs a range of elicitation techniques (some examples will be considered in chapter four).

In a separate analysis of 20 similarly-structured oral interviews recorded on video and audiotape, Lazaraton concluded

the overall structural organization of conversation seems to be operative in these encounters ... In contrast, the system of turn-taking which is at work in these encounters is not the 'locally managed' one of conversation, but a pre-specified system which defines an interaction as an instance of 'interview'. (Lazaraton, 1992: 383)

Such highly structured interviews may improve reliability through elicitation of more consistent language samples, but at the same time they pose a threat to validity through reduced authenticity (Underhill, 1982; Riggenbach in Young and He, 1998: 55)

The use of normal conversation as the yardstick for comparison also raises questions about the transferability of these findings to contexts where the target language use can be clearly described. The work sample approach to performance assessment (McNamara, 1996) is premised on the determinability of target language contexts of use, and it is against these that the interaction of the test should be compared. It seems likely that the structure of institutional discourse in an occupational or professional environment will be much more like the interview test than normal conversation where each participant enjoys equal rights and equal responsibilities (Drew and Heritage, 1992).

Ross and Berwick (1992: 160) described the standard OPI format as "a hybrid of interview and conversational interaction". Moder and Halleck (in Young and He, 1998) compared the OPI against the job interview, as a more realistic yardstick than normal conversation:

For many non-native speakers, the majority of their interactions in English may occur in contexts where there is an institutionally based power discrepancy between participants and where the interaction is purposive. (Young and He, 1998: 118)

They found that native speakers in job interviews took more shorter turns and asked many more questions as non-native speakers in language proficiency interviews, and suggested this was because the native speakers felt less need to display language use for its own sake and were more comfortable checking the background knowledge of the interviewer in order to help meet the maxims of quality and relevance. Overall, they concluded that the general interview was a very suitable frame within which to view the OPI, and it was clearly less comparable with informal conversation.

A further practical difficulty for test validation is that interaction of any kind is arguably 'co-constructed' by all the participants in it, and that therefore "interactional competence is not an attribute of an individual participant" (He and Young in Young and He, 1998: 7). This is in contrast to communicative competence, which as a construct was clearly conceived, for example by Canale and Swain (1981) as a trait or bundle or traits that can be ascribed to and assessed in an individual.

Even if there is only one candidate and one interlocutor involved as participants in an oral test, they share responsibility for the success, or the failure, of the interaction. Any reluctance on the part of the interviewer to be fully involved may undermine the interaction and thus the ability of the candidate to fully display their conversational skills. For the analysis of the Five Star test, interaction was omitted from the list of skills that expert panel members used to analyse test tasks and events, in part because of the difficulty of assigning the success of any given interaction to a single participant in it (see section 5.1.1 below).

Nonetheless, the working definition of interaction used by the Five Star test development was restricted to the individual candidate: "a learner's ability to facilitate participation in a one-to-one discussion through the employment of negotiation devices such as confirming understanding, requesting repetition and seeking clarification" (Pollard, 1997). The definition included 'one-to-one' because the population profiling analysis of the target language use domain described in 4.2 below identified one-to-one encounters as the most common format of NS-NNS (native- to non-native speaker) events. This operationalisation of the construct as an individual variable led directly to the framework used for the analysis of verbal interaction strategies by the expert panel, described in 5.1.1 below.

Given the complexity of the interaction construct, it is not surprising that there is no widely agreed definition, and the kinds of features that occur are operationalised in different combinations in the rating criteria for different tests. The University of Michigan's Examination for the Certificate of Proficiency in English, for example, has a direct speaking component that is actually called Interactive Oral Communication, and among the salient features that assessors are to look for are 'interactional facility' and 'sensitivity to cultural referents'.

The behaviour to be evaluated under 'interactional facility' is "monitors interaction, seeks clarification when appropriate, takes turn at appropriate time, properly engaged, appropriate eye contact/posture" while 'sensitivity to cultural referents' includes "establishes common frame of reference, initiates clarification, rephrasing, concrete relevant examples". (ELI, 1999). These two criteria overlap with the Five Star definition

above, but interestingly they also explicitly include turn-taking and para-linguistic behaviour.

3.3.3 Face validity

The application of MTMM was welcomed as a possible route to allow for the statistical validation of the kind of subjective tests that were coming back into fashion with the communicative approach, such as direct oral tests.

Unfortunately, the greatest appeal of oral interviews to the technically untrained language teacher rests with their high face validity. This appeal tends to cloud and confuse the need to validate these tests. As a result, although oral interviews are becoming more and more popular among language teachers and testers, this popularity far outruns any technically demonstrated validation... (Stevenson, 1981: 37)

The fact that their acknowledged 'high face validity... tends to cloud and confuse the need to validate these tests' encapsulates that author's view, widely shared by professional colleagues, that face validity is inferior and superficial, because of its lack of technical (i.e. statistical) foundation.

It forms the major claim for the validity of oral interview measures... yet is not recognized by the measurement tradition as having any bearing on a technical consideration of what a test measures. Rather, face validity can be considered to be appearance of validity in the eyes of the metrically-naïve observer. (page 41)

Presumably, the test-taker, whose conscientious participation in the test event is taken for granted but strongly influenced by perceptions of test validity, is included in the rather patronising label 'metrically-naïve'. It raises the question whether the inclusion of live oral interaction in a test of communicative competence can be justified by appeal to face validity only; to content validity, on the basis of a skills approach that says that writing should be tested by writing tasks, speaking by speaking tasks, and so on; or to

construct validity, on the basis that the theory of communicative competence requires strategic and pragmatic competence to be tested in direct ways that indirect tests fail to do.

In the empirical tradition, exemplified by Stevenson above, face validity was never seriously considered as an aspect of validity. Bachman (1990) charts its decline in a section entitled 'post mortem: face validity', noting that the American Psychological Association (APA 1974) standards stated that face validity was not an acceptable basis for interpretive inferences from test scores, while its final interment was marked by its total absence from the APA (1985) edition. However, the appearance of a test clearly has a significant effect on its acceptability to stakeholders, and "the 'bottom line' ... is whether test takers will take the test seriously enough to try their best, and whether test users will accept the test and find it useful. " (Bachman, 1990: 288)

The advent of communicative testing led to a higher value placed on face validity in the rationalist tradition than in the empirical/psychometric one.

In the past, face validity was regarded by many test writers simply as a public relations exercise. Today, however, most designers of communicative tests regard face validity as the most important of all types of test validity. (Heaton, 1988: 160)

Heaton goes on to associate face validity in communicative testing with the appearance of authenticity. Thus authentic source material for test tasks, such as news bulletins, promotional materials or newspaper articles must look as much like the real thing as possible, even if they are in fact only semi-authentic and have been specially written for the purpose of the test.

Others have taken a view on face validity between these two extremes, describing face validity as necessary but not a substitute for other forms of validity (Weir, 1990: 26).

Hughes is more apologetic but also uses the word 'important':

Face validity is hardly a scientific concept, yet it is very important. A test which does not have face validity may not be accepted by candidates ...[whose] reaction to it may mean that they do not perform on it in a way that truly reflects their ability' (Hughes, 1989: 27)

Another author considered this a particular issue for testing adults: "Especially in adult testing, it is not sufficient for a test to be objectively valid. It also needs face validity to function effectively in practical situations" (Anastasi, 1982: 136) but she does not explain why this argument is less applicable to children.

Face validity can be quantified empirically by the simple expedient of asking test participants their views. For example, Huhta and Randell (1996) asked students to evaluate different task types with the question 'How well do you think this test measured your English reading comprehension?' on a 1 to 5 scale (in this case, two open-ended test formats scored slightly higher than multiple-choice type tasks, but overall the face validity ratings clustered around the mid-point).

3.3.4 The empirical validation of oral tests

Statistical correlation between oral and written tests have often been used to justify the use of the latter in preference to the former. Adherence to a global rather than a discrete-point approach, and a considerable faith in the meaning of correlation, allowed Streiff to claim "significant correlations were found between oral and written cloze scores, and they were substantial enough so that for a bilingual population Written cloze could

be substituted for oral cloze as a measure of oral language proficiency" (Streiff in Oller and Perkins, 1978: 94) The correlation coefficients used to justify this claim were actually quite low, ranging from 0.54 to 0.81. Squaring these figures to produce more meaningful coefficients of common variance (Crocker and Algina, 1986: 35) indicates a shared variance between the tests of one to two thirds only.

The distinction between norm-referenced tests (NRTs) and criterion-referenced tests (CRTs) is for Brown (1996) critical in determining the value of correlation studies. Unlike NRTs, which are by definition designed to produce normally-distributed scores, CRTs are not and in many cases will produce skewed results, and the assumption of normal distribution which underlies correlation analysis will be violated (Brown, 1996: 232).

Oller devoted a chapter to tests of oral productive communication, and felt there was a serious need for better oral tests: "here, more than elsewhere, ... we are forced into largely unresearched territory". (1979: 308) He reviews two oral testing procedures in widespread use at that time - the Bilingual Syntax Measure and the Ilyin Oral Interview, both of which are based around cartoon pictures and stories, and suggests how they could be adapted to meet his pragmatic naturalness criteria better. He criticises a third, the Oral Communication Test associated with the University of Michigan, for lack of continuity between items: "the Oral Communication Test violates the pragmatic naturalness of constraint of meaningful sequence. That is, there is no temporal connection of the sort characteristic of normal discourse between consecutive items on the test" and again suggests how the test could be modified.

Oller then reviews the Foreign Service Interview, as the standard oral interview procedure in use, and notes that while the rating scales such as accent, grammar and vocabulary are based on discrete point theory, a factor analysis does not suggest that they contribute differently to the test variance, and concludes that the "utility of the FSI procedure is dependent primarily on the ability of raters to differentiate performances on one basic dimension - the pragmatic effectiveness of speech acts" (1979: 325). He also considers how other oral tests procedures, such as reading aloud, oral types of cloze test, and various narrative task, can be considered pragmatic.

The emergence of the communicative testing paradigm combined with a wider range of statistical techniques gave a new impetus to attempts to reconcile the content validity of direct oral tests with the perceived superiority of empirical validation. An important collection of papers delivered at a colloquium in 1979 was subsequently published as *The construct validation of tests of communicative competence* (Palmer et al., 1981) and this set the agenda for the north American school for the 1980s.

Some of the papers focused specifically on the validation of oral testing procedures, and five of these are reviewed briefly here.

- a) In an analysis of around 60 contemporary oral tests, Madsen and Jones concluded that the great majority included sub-tests and multiple elicitation techniques, rather than single procedures, and that "without abandoning their interest in integrative examinations, test makers evidence a strong interest in approaches that are quantifiable (e.g. number of responses in 30 seconds, exact word criteria in elicited imitation, readily-identifiable answers to picture-cued questions)". They consider a

range of elicitation techniques on a spectrum from communicative to discrete item, but generally favour an analytic view of what is to be tested, expressing concern about the content validity of oral tests from a sampling point of view: "many important linguistic structures are not produced. Because the language is random, it is not a good sample of what is taught in the classroom or what is considered to be a minimal standard of proficiency at any particular level". (16) While acknowledging that "a rising interest in communicative competence has forced us to examine more closely what else besides linguistic facility contributes to effective communication in a second language" they conclude that "unfortunately, these additional features pose very difficult problems for testing". (Madsen and Jones in Palmer et al., 1981: 21)

- b) Using as a yardstick an established oral test rating scale, the U.S Department of State's Foreign Service Institute (FSI) criteria, Lowe focused on tasks and question types which promoted content validity, defined as "the degree to which the oral interview procedure makes possible the elicitation of a speech sample evaluable in terms of the FSI criteria" (Lowe in Palmer et al., 1981: 71). Lowe considered that content validity of an oral interview could be strengthened by using a sequence of four phases, warm up, level check, probes and wind-up. The warm-up and wind-up do not contribute significantly to the evaluation; the level check is the main instrument for assessment, with probes to test for evidence of a higher level of proficiency. The interviewer checks off a testing protocol of tasks, language functions and question types which are considered characteristic of the speaking behaviour at each of the eleven levels of the test. For example, 'checked for minimum courtesy requirements' is at level 1, satisfying 'routine social demands' is

at level 2, answering hypothetical questions at level 3, and so on. "Properly filled out, the protocol should have most (or all) of the boxes checked at the candidate's proficiency level, with some boxes checked at the next higher level in order to make sure that probes have been attempted" (page 76) This model can be seen as a precursor of an adaptive test; the interviewer pushes the candidate up the scale of tasks until he or she falters, then retreats a little for the wind-up phase, in order to give the candidate a 'feeling of accomplishment'. Thus one administration of the test for a candidate of lower proficiency might end where the next is just beginning to reach a better candidate's true level of performance.

- c) Another well-established north American oral test is the Ilyin Oral Test, consisting entirely of a series of cartoon pictures of routine events and activities in a person's day, which act as prompts for scripted interviewer questions. Engelskirchen and colleagues examined the reliability and validity of the test by analysing the inter-rater agreement, arguing that "the sort of agreement that is required across native judges is not only a prerequisite reliability criterion, but is actually the most appropriate validity criterion for any such test." (Engelskirchen, Cottrell and Oller in Palmer et al., 1981: 84) No theoretical justification is offered for this statement. Their principal instrument was a factor analysis of scores awarded to taped interviews of the same group of subjects by 20 different raters.

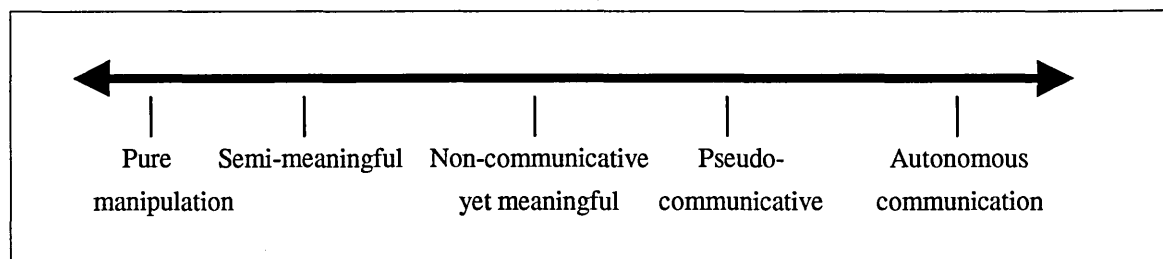
A single principal component of the scores emerged as the panel consensus, and the loading of each rater with that consensus was interpreted as the validity coefficient for each judge on their own (correlations ranging from $r=.60$ to $r=.95$), and their overall agreement with it as an index of overall validity (multiple correlation .81,

accounting for a total variance of .66) As a form of content validation, panel raters were asked to comment on the relationship between the pictures presented and the questions to be asked (which being entirely scripted leave no room for interviewer variation). The two specific questions asked about the degree of fit between the picture and the question, and the naturalness of the question itself; some being much more natural than others: 'what time does Bill usually study?' cf. 'Ask a question about this picture with the word "if"'. By comparing judges' ratings of each test question with a conventional analysis of item discrimination, the authors concluded that items scoring higher in naturalness and picture fit also discriminated better, and therefore that "overall discrimination might be improved significantly by making the items conform more closely to the pragmatic requirements of communication ... we are encouraged to believe that oral tests of this sort can be refined to extremely high levels of reliability and validity". (page 92) However an apparent failure to operationalise 'naturalness' undermines its equation with communicative criteria, and the use of the same judges to rate the items for naturalness after using them to score tests undermines the independence of the two process.

- d) Shohamy carried out a study to investigate the concurrent validity of an oral interview based on the FSI with a written cloze test, using Hebrew as the target language. Correlations of around .85 were reported, which "may be related to the instruction and teaching methods used". There is no discussion of why an open-ended oral test and a written gapfill test should correlate highly, given that they are testing such different skills; however, both are pragmatic tests of language, in Oller's terminology, and this may be the underlying unspoken paradigm (Shohamy in Palmer et al., 1981).

e) Another imaginative attempt to marry the desire for direct oral test with the need for reliable systems of scoring was a pair of picture description techniques called COMTEST and PROTEST. These consisted of a candidate describing one of four similar pictures until sufficient detail enabled the examiner to identify which was being described (PROTEST); and for the candidate to ask the examiner a series of questions to determine which of the four pictures the examiner has in mind (COMTEST). IN both cases, the score is the amount of time needed to complete the task. In a study of concurrent validity, PROTEST and COMTEST correlated at .62 with each other, but only at .45 and .34 respectively with an oral interview, which might be expected to show high concurrent validity. Moreover, comparison of the interview against a separate dictation test yielded a correlation of .70, which ought to have been lower on a convergent/divergent skills construct.

Palmer suggested that PROTEST and COMTEST required very specific and restricted kinds of speech behaviour, such as a limited range of speech acts (Searle 1969) and a lack of freedom to vary the topic. Conversely, deductive reasoning ability and testwiseness (learning effect) were considered to be unusually important influences in successful performance on these tests. Therefore, the oral interview was testing something different from COMTEST and PROTEST, and high concurrent correlations were not necessarily to be expected. This difference of language construct was modelled on a scale which shows language production as varying from pure manipulation at one extreme to autonomous communication at the other, illustrated in Figure 4.

Figure 4**Language production scale (Palmer, 1981)**

Palmer's analysis of the kinds of speech behaviour required identified PROTEST and COMTEST as generating 'pseudo-communication' on this scale, and therefore offering low concurrent validity with the oral interview, 'the most widely accepted type or oral proficiency test'. On the brighter side, Palmer concluded, they were quick and easy to administer and required minimal training for examiners. The reliability/validity trade-off model (Underhill, 1982) would see these as two sides of the same coin; the act of standardising the test format in order to generate fair and comparable language samples on a single objective scale (time taken) itself constrains the validity, in a broad sense, of the spoken language produced. (Palmer in Palmer et al., 1981: 138)

Hughes (1989) suggested a sampling approach to identifying the tasks that should be tested for both speaking and writing. "The basic problem in testing oral ability ... is to set tasks that form a representative sample of the population of oral tasks that we expect candidates to be able to perform' and the first step in this process is 'specifying all appropriate tasks", using a framework of operations (language functions or speech acts such as narrating, eliciting), types of text (dialogue, telephone), addressees and topics. It is a simplified form of the deterministic needs analysis model, which assumes that it is

in fact possible both to define the population of oral tasks for a particular learner or group of learners, and to describe the language they will need in a comprehensive way. It assumes that language use is predictable.

Statistical correlation can also be used to claim criterion validity for pen-and-paper tests as tests of spoken English. A report comparing the TOEIC-LC (Test of English for International Communication), a multiple choice pen-and-paper test with a listening comprehension component, with the Language Proficiency Interview, a direct oral test, concludes

Although the LPI requires a response in spoken English, while the TOEIC requires an examinee to answer questions printed in English in the test booklet, both the LPI and TOEIC-LC measure the underlying ability to comprehend spoken English. The fact that correlations proved to be consistently high between LPI and TOEIC-LC strongly suggests that both tests are, in fact, measuring the common ability to understand and use spoken English (Wilson, 1993: 11)

In fact, the actual correlations reported, for a sample of nearly four hundred candidates from four countries, were of the order of .73 to .76 (Wilson, 1993: 9). Using the coefficients of common variance, obtained by squaring the correlations (Crocker and Algina, 1986: 35) as a measure of the amount of overlap in the variances between the scores on the two tests suggests that they have barely 50% of their variance in common.

3.4 Adaptive testing

An adaptive test is one in which responses to earlier items influence the selection of later items in each test administration, so that two consecutive subjects taking the same test actually face few or none of the same tasks or items. The strongest argument for it is efficiency:

A critical problem facing a mass-administered test is that it must be assumed that there is a relatively broad range of ability to be tested... Most examinees' abilities seem to lie in the middle of the continuum. Thus, mass tests match this by having most of their items of moderate difficulty with fewer items at the extremes. The consequence of this ... has historically been that the most proficient examinees have had to wade through substantial numbers of too easy items before reaching any that provided substantial amounts of information about their ability' (Wainer et al., 1990: 9-10)

Weiss (1982) describes this as a 'bandwidth-fidelity dilemma'. He describes research that demonstrates how adaptive tests based on item response theory can provide a solution to this dilemma, by permitting "measurements of equal precision throughout the range of the trait being measured while maintaining high levels of efficiency" (page 474)

While efficiency is a criterion that is strictly independent of validity, tests that are grossly inefficient can be seen as suffering poor construct validity. Messick (1989) identified two types of 'surplus construct irrelevancy', where test variance does not originate from the construct under measurement, either because the task is too difficult or too easy. Adaptive tests increase validity by tailoring the test to the students' level of performance (Laurier, 1996)

In addition to increased efficiency, Wainer et al. note such other benefits of adaptive testing as test security; immediacy of scoring; greater ease of altering test contents and trialing new items; and individuals being able to work at their own pace.

Some of these benefits are the direct result of computer delivery, which is closely associated with adaptive testing, but not all adaptive tests are computer-based (a live oral interview, for example) and many computer-based tests are not adaptive. Weiss

(1982) describes Binet's intelligence test as the first adaptive test. It was administered by a trained psychologist, who used an item selection rule to select subsequent items based on the examinee's responses to items already administered. "Because individual [live] adaptive test administration is expensive, and paper-and-pencil adaptive test administration is both cumbersome and inefficient, most current adaptive tests are administered by computers". (Weiss, 1982: 474).

Even where adaptive tests are computer based, they may not exploit the full potential of the medium: "...the biggest limitation of CATs [computer-adaptive tests] is that test developers simply treat CATs as an alternative delivery format (albeit with certain advantages), instead of as a new technology that can be used to expand our thinking of item and task types". (Deville, 2000)

Weiss quotes research results which 'showed adaptive tests requiring half the number of items as that of conventional tests to achieve equal levels of reliability, and almost one-third the number to achieve equal levels of validity' estimated by concurrent correlation against a criterion test (Weiss, 1982: 473). Laurier (1996) compared computer adaptive tests (CATs) with pencil and paper tests (PPTs) and found the CATs as reliable as PPTs and generally 50% shorter for placement purposes, with CATs particularly effective at the upper and lower extremes of the range of measurement. Recent research on the application of item-banking to the development of the new Communicat computer adaptive test (UCLES, current) has shown that 18-20 items were sufficient to identify a candidate's level of general proficiency in English, to a pre-defined estimate of error (Corcoran and Jones, 1999; Williams, 2000)

The potential of adaptive testing of language, through the delivery systems made available by the closely associated computer technology, has only recently begun to be exploited. However, those oral tests of language proficiency that are direct and 'live' - as opposed to recorded, or indirect tests claiming to reflect oral proficiency by concurrent correlation - have for a long time exploited the fact that the interlocutor can alter the level of difficulty of tasks or questions posed to match more closely the level of the subject, and indeed this is in a sense a reflection of what we do in everyday conversation when we cannot be heard or understood properly.

The basic notion of an adaptive test is to mimic automatically what a wise examiner would do. Specifically, if an examiner asked a question that turned out to be too difficult for the examinee, the next question asked would be considerably easier. (Wainer et al., 1990: 10)

As well as consciously selecting easier tasks, oral examiners may adapt the procedure in response to candidate responses in other ways. In research to investigate the different kinds of verbal support given by interlocutors, Lazaraton found that "interviewers routinely modified their question prompts ... in response to perceived trouble for the candidates by reCompleting question turns, by suggesting alternatives to choices presented and by reformulating the questions altogether." (Lazaraton, 1996: 153)

In tasks where these interviewer modifications may be evident to the candidate, they would qualify as 'reciprocal' for Bachman and Palmer (1996), who contrast adaptive and reciprocal tasks. Reciprocal tasks give the candidate "feedback on the relevance and correctness of the response, and the response in turn affects the input that is subsequently provided by the interlocutor" (1996: 55) whereas an adaptive task only requires that the response affects subsequent input. Thus reciprocal tasks must be adaptive, but adaptive tasks need not be reciprocal. There is some overlap between

Bachman and Palmer's use of the terms 'reciprocal' and 'interactive'; this is discussed in 3.6 below.

The issue of interlocutor feedback on the appropriateness of a candidate's response is complicated by the dual role that the interviewer may be playing of interlocutor and assessor combined. Van Lier gives an extreme example of an interviewer who asks a 6-year old child where his/her mother is and what she does; the candidate replies 'She's dead' and the interviewer's response is 'Ah - she's dead. Very good'. This is positive reciprocal feedback to the language content of the candidate response, indicating that it is a wholly well-formed and appropriate reply to the question asked. Unfortunately, the interviewer's response is itself wholly inappropriate by normal conversational and ethical criteria (van Lier, 1989: 499).

Many universities in the USA are currently using computer adaptive tests, primarily for the purposes of placement into language classes. Universities there have been in the forefront of testing research and development, and in addition, they already have the computer laboratories installed and so introducing CATs thus does not incur additional hardware costs. With large pre-sessional programmes, placement tests that are computer-administered and marked are labour saving and, since the candidates only see a placement test once, item pools can be smaller and test security less contentious than for other testing purposes (Deville, 2000)

Adaptive tests typically follow one of two systems for the selection and sequencing of items. Either the routes through the test are pre-programmed, like a branching programmed learning, and although there may be a very large number of permutations,

the selection of each task is predictable from the outcome of the previous task. The test ends when the algorithm comes to the end of a route. The Five Star test is an example of such an algorithm. Alternatively, items may be flagged by difficulty level and stored in an item bank from which they are selected at random when the algorithm has determined that an item of a particular level of difficulty is required. As the test proceeds, an estimate of the candidate's ability is progressively refined, and the test ends when the standard error of this estimate falls within pre-defined boundaries. Large-scale adaptive tests follow this pattern, but require a very large item bank to operate with precision and test security (Hill, 1995).

Some specific examples of computer adaptive tests are considered in the next chapter, by comparison with the Five Star test.

3.5 Item response theory (IRT)

The classical model of mental measurement posits an observed score consisting of two components, a true score of the trait being measured, and an error, which may in turn derive from a number of different sources and types. This contrasts with other models of test response, and particularly types of scales based on item characteristic curves. Like conventional item analysis procedures, these allow individual items in a test to be described in terms of their difficulty level and discriminability, but unlike the conventional models, these item qualities can be assessed independently of the sample on which they were based.

Classical test theory rests upon the assumption of a general linear model which is basic to correlational analysis, linear analysis and factor analysis. It is concerned, therefore, with additive errors of measurement and score components. Item response theory [IRT], on the other hand, is couched in stochastic terms, with a probabilistic response model of which the parameters express certain item characteristics. (Kline, 1993: 67)

The seminal work on item-response theory was *Best test design* (Wright and Stone, 1979). The authors acknowledge the Danish mathematician Georg Rasch, working on intelligence tests in the 1950s, as the person who first produced a stochastic model for defining item difficulty independently of the candidate sample, and candidate ability independently of the item sample (Rasch, 1960).

In essence, the model treats both the candidates and the items as samples of larger populations, and produces estimates of candidate ability and item difficulty, expressed on the same scale, in terms of the likelihood of a candidate of ability β being able to get an item of difficulty δ right. This likelihood is actually expressed as a logarithm of the constant e , and the unit of measurement that results is generally known as the *logit*, as a contraction of 'log odds units' (McNamara, 1996: 165). For each item and each candidate score, the model also produces an estimate of the error and an estimate of the degree of goodness of fit to the model.

The whole approach is still often referred to eponymously as Rasch analysis, but in fact it has grown and diversified enormously as a family of related analyses, and the more general term item-response theory (IRT) is used here. A major distinction reflected in the literature distinguishes one-, two- and three-parameter models. Rasch analysis, in the narrow sense, is a one-parameter model that considers item difficulty; a two-parameter IRT model adds a parameter for discrimination, and the three-parameter model a further parameter for guessing.

Although two- and three-parameter models are more sophisticated and produce item estimates for each parameter, they suffer from three disadvantages: they are generally less robust, they require much larger datasets to generate meaningful information, and they are effectively restricted to the analysis of dichotomous data and cannot be used on partial credit data such as the Five Star test (McNamara, 1996: 259). The particular version used for data analysis in this research is therefore the one-parameter partial credit model, and this is explored in more detail in section 5.2.1.

“The essential feature of an IRT approach is that a relationship is specified between observable performance on test items and the unobservable characteristics or abilities assumed to underlie this performance... The characteristic measured by a given set of items, whether a psychological attribute, a skill, or some aspect of educational achievement, is conceived of as an underlying continuum, often referred to as a latent trait or variable... This underlying continuum is represented by a numerical scale, upon which a person's standing can be estimated using his/her responses to suitable test items... Items measuring the trait are seen as being located on the same scale, according to the trait level they require of testees.” (Baker, 1997: 19-20)

The logit scale is moreover an interval scale, allowing direct comparison between item difficulty scores, or between candidate ability scores, or a longitudinal plotting of an individual's scores over a period of time. The measures of item difficulty are conventionally distributed about a mean of zero, with a negative value signifying an easy item and a positive value a difficult item. Being measured on the same scale, a negative value for candidate ability indicates low ability and a positive value a subject with a high level of ability. Each estimate is accompanied by an error term.

'Although IRT has many obvious advantages, its real strength was that it could deal with items one at a time. It posited an underlying, unobserved trait, on which the items were linearly arrayed from the easiest to the hardest. The goal of testing was to be able to array the examinees on the same continuum as the items, from novice to expert. This goal meant that one did not have to present all items to all individuals, only enough items to allow us to accurately situate an examinee on the latent continuum. The power to do this did not exist comfortably within the confines of traditional true score theory and yet was a natural outgrowth of IRT. In fact, the capacity to rank all examinees on the same continuum, *even if they had not been presented any items in common*, gave rise to the possibility of a test that was individually tailored to each examinee. Such a test is called Adaptive, and many believe that adaptive testing is the *raison d'être* of IRT' (Wainer et al., 1990: 9)

Of a number of features that distinguish IRT from classical test statistics, the crucial advantage here is the ability to compare candidates on the basis of their performance on different samples of items, and items on the basis of how well different samples of candidates did on them. This makes it particularly suitable for the analysis of data from a test such as Five Star. Conventional item statistics such as item facility and item discrimination, and internal reliability measures such as the Kuder-Richardson formulae known as KR20 and KR21, can only be used where there is a complete dataset, with all candidates having attempted all items. These statistical techniques are therefore unavailable for an adaptive test such as Five Star.

Weiss identifies four other advantages of IRT-based adaptive testing.

1. Because person ability estimates and item difficulty estimates are based on the same scale, selecting an item of appropriate difficulty is much easier than if these two variables were based on different yardsticks.
2. It is possible to estimate ability levels using any subset of the item bank, so that different items can be deliberately chosen for different individuals, but their ability scores still compared directly.
3. It is not necessary to constrain the structured branching of the algorithm in advance; so long as the items in the bank are all tagged for difficulty (and potentially for other parameters also), the computer can simply search the pool for an item that meets the relevant criteria. The Five star test does not do this, but rather has a pre-defined branching algorithm.
4. It offers an objective criterion for deciding when the test is to end:

termination of an adaptive test can be based on the precision of the measurements obtained. IRT scoring procedures make it not only possible to estimate ability levels after each item is administered and answered but also make it possible to determine the precision (standard error) of each ability estimate. These standard errors can then be used as criteria for terminating the adaptive test. (Weiss, 1982: 475-6)

Assumptions

Two statistical assumptions made by IRT relate to unidimensionality and local independence (Baker, 1997: 30; Crocker and Algina, 1986: 342). Unidimensionality requires that all the items or tasks in a test be measuring the same trait or combination of traits. The construct underlying the test need not be psychologically simple, but the items should function in the same way, in other words, completing the tasks should call on the same skills or abilities. Local independence requires that performance on any one item is not influenced by success or failure on any other item. Both of these assumptions have implications for the Five star test, and are considered further in section 5.2.1.

Item response theory in language testing

Although the validity of IRT was a contentious issue and not unanimously accepted at first (McNamara, 1996: 5), language testing has come to welcome it as:

... a useful additional tool for the test constructor. It can be used for identifying items which do not fit into a test or for identifying students who do not fit in with their testing group. It is useful for detecting bias, and can be used for the analysis of the results of subjective as well as objective tests. It is also invaluable for computer-adaptive testing. (Alderson et al., 1995: 91-92)

The potential contribution of IRT in language testing is particularly great when trying to use data collection from one sample to validate a test for use on a much greater population, but traditionally being unable to separate the influence of the test from the influence of the sample:

The results of analyses carried out using ... classical test analysis procedures have one major drawback. The examinees' characteristics and the test characteristics cannot be separated, so that the results of the analyses are only true for the actual sample on which the trials were carried out. The results will not apply to samples of students at different levels of proficiency. It is not, therefore, possible to provide any fixed measure of a test's difficulty... Measurement using Item Response Theory is designed to cope with this problem. We can use it to develop an item difficulty scale that is independent of the sample on which the items were tested... (Alderson et al., 1995: 89-90)

The original model proposed by Rasch was for the analysis of dichotomously-scored data, where candidate responses to any item or task are categorised as either 'right' or 'wrong'. In language tests such tasks are typically multiple-choice or true/false items, but there are a number of other more or less objectively-scored task types in common use, such as cloze tests (whole or part-word gapfills), word order, matching exercises and so on. The Five Star test, however, is in a different category, as there are three possible score outcomes to each task, and a variation of the basic item response theory model is needed to take account of the 'partial credit' data. This is considered in more detail in chapter five below.

Some examples of the use of IRT for validation in language teaching:

- a) Evaluating raters' interpretation and range of use of rating sub-scales (e.g. grammar, vocabulary, pronunciation) (Milanovic et al., 1996);
- b) Estimating the optimal length of a computer-adaptive test (Laurier, 1996)
- c) Identifying individuals with highly idiosyncratic patterns of performance, who might simply score badly on a conventional test because they deviate significantly from the population norm on which it is based (Masters in De Jong and Stevenson, 1990)
- d) Exploring the validity of the Speaking Proficiency Guidelines on task difficulty posited by the ACTFL (American Council on the Teaching of Foreign Languages), Stansfield and Kenyon (1996) asked 700 modern language teachers in public

schools in Texas to scale 38 different speaking tasks by difficulty, in terms of the level of ability required to perform each one, and compared the outcome with the ACTFL guidelines.

- e) Validating panelists' ratings of the relative difficulty levels of reading subskills (Lumley, 1993, described in more detail in section 5.1.1 below)

3.6 Summary

Language proficiency is generally seen as a psychological construct and the measurement of language draws heavily on the literature and assumptions of psychometric measurement. The validation of language tests therefore shares many common concerns with psychological measurement, such as the difficulty of satisfactorily operationalising ultimately unobservable constructs; lack of agreement over the measurement tools; lack of well-defined units on an agreed measurement scale; and uncertainty over the size of the error component in any measurement.

Some of the tensions in the field of language testing therefore reflect debate in the field of psychological measurement and language testing research, at the formal, academic end of the continuum of activity, is dominated by the principles and practice of psychological measurement as codified by the American Psychological Association. Thus validity is currently seen as a unified concept drawing on diverse sources of evidence, but with construct validity in particular as the 'first among equals'. The need in the academic community to maintain a scientific rigour in the endeavour of language measurement exerts a constant pressure towards quantitative methods and an empirical

approach both to measurement and to the validation of that measurement. In its extreme form, this is manifested in an exclusive focus on *post-hoc* empirical validation to the exclusion of *a priori* content specification and in the search for construct validation through purely statistical means.

Among the forces for inertia has been the use of classical test statistics which require complete data sets and generate statistics that are sample dependent, which can only be extrapolated with great caution to other samples or a wider population. Item-response theory (IRT) approaches offer some escape from these constraints and were first elaborated in the early 1960s, but only in the last 20 years has the lower cost and increasing availability of micro-computers brought IRT within the grasp of a wider circle of testing and measurement specialists. The same process has made feasible for the first time the use of computer-based tests on a large scale, and the combination of widely-available computer platforms with IRT analysis makes fertile ground for the development of adaptive tests. By and large, however, the test instruments themselves remain anchored in the analytic approach to language measurement.

In this paradigm, language performance testing in general and the assessment of spoken language in particular is problematic in that it introduces a new set of variables and possible sources of error. Although the communicative methodology may be largely accepted as the dominant paradigm in language testing, many widely used and respected tests do not conform to it and there is considerable academic debate over the identity and status of its major tenets.

The literature of language testing may be dominated by such a paradigm but world-wide language testing is not an exclusively or even largely an academic research activity. The great majority of language testing goes on every day in schools, colleges and other institutions around the world, carried out by teachers who may be well-qualified and experienced as teachers but who have little or no formal training in testing and assessment.

Pen-and-paper tests of knowledge of language are well-established, form part of every language learning programme evaluation and require no great expertise to administer or mark. The greater need is to develop tests of communicative performance that are face valid through obvious parallels with real-world language use in everyday contexts and so command the respect of teachers, learners and other stakeholders, yet at the same time can be validated with sufficient theoretical and empirical rigour.

The combination of computer resourcing and adaptive testing analysed by IRT statistics now offers opportunities to explore how communicative performance testing can be realised in small- or medium-scale contexts and chapter four turns to look in more detail at one such test, the Five Star test, that combines some of these aspects.

Chapter 4 The instrument: the Five Star test and comparisons with other tests

4.0 Introduction

4.1 Background

4.2 Five Star test description

4.3 Summary of key communicative features

4.4 Comparison with other current tests of speaking

4.5 Comparison with other computer-based tests

4.6 Summary

4.0 Introduction

This chapter introduces the Five Star test in more detail. It describes the origins and development within a specific geographical and cultural context and summarises the features of the test against aspects of the communicative methodology that were discussed in chapters 2 and 3. The Five Star test is compared with other currently available test of speaking and with other computer-based tests.

4.1 Background

British Aerospace (BAe) has for thirty years been a major contractor to the Saudi Arabian defence agencies, providing logistical and training support as well as military hardware. The recruitment and personnel function of this relationship includes a large English language training programme, delivered at military bases all over the country. The American Language Course (ALC) and its associated tests described in chapter two as an example of the structuralist approach was and in many cases still is being used as the core of the curriculum, despite mounting professional concern (e.g. Weir, 1990; McNamara, 1996) that such indirect, multiple-choice tests are not appropriate on their own and have a negative washback effect on teaching (Al-Ghamdi, 1994, quoted in section 2.4 above).

In 1992 British Aerospace established an offset company in Saudi Arabia, Saudi Development and Training (SDT), to transfer British training practices to Saudi Arabia and make them available locally as part of a joint Saudi and British commercial enterprise. This created the potential backing for developing products to meet training needs that were not already catered for. One such need was identified as an English language proficiency test particularly suitable for companies to screen young adult applicants, in the absence of a commercially available general language proficiency instrument in post-high school training departments and institutions in Saudi Arabia. Existing English language tests fell into two categories: either they were indigenous to the Saudi education system, in secondary or tertiary institutions, or they were external examinations from Britain or America designed to test general/academic English proficiency. In either case, they were unsuitable for Saudi nationals applying for work

positions and training programmes. "It is difficult to identify a commercially available test that could be used in post-high school training departments and institutions in Saudi Arabia, and there are powerful arguments for the local development of tests in contexts where the resources can bear it." (Pollard and Underhill, 1996: 49)

The Saudi labour market

The broader socio-economic context is one in which the government which has traditionally been the only large-scale employer of its own citizens is seeking to transfer some of these responsibilities and costs to the private sector, and a policy of 'Saudisation' is used to encourage this transition. In essence, this is a process of skill transfer from expatriate workers to Saudi nationals, gradually restricting the access of the former to the Saudi labour market while training up the latter to take their place. The policy is driven by long-term strategy against an economic background of uncertainty and fluctuation in the price of oil, a high rate of population growth, an urgent need to diversify sources of income and wealth creation, and a desire to rein in the sometimes extravagant expenditure of the boom years. In addition, one of the legacies of the Gulf War was a huge national debt at a time of declining oil revenues.

Robinson (1996) describes the background in a review of the Saudi labour market:

The massive transfer of wealth to the region enabled massive national development plans to be put in place. To execute these plans there were simply not enough Saudis to go round and so labour and skills were imported into the region... Infrastructure in the form of electrification, health care and pure water triggered a decline in infant mortality and a baby boom... It is this baby boom that is now giving the governments of the region pause for thought as to how to incorporate these many young men and women into the work force" (Robinson, 1996: 2).

The rapid establishment of this infrastructure under national ministries or institutions led to an imbalance in the labour market:

during the early years of national development plans ... the nationals emerging from the nation's schools and universities tended to be snapped up by the growth of the government and civil service or into large parastatal companies ... at the end of 1415AH¹ that 88% of Saudi employees work in the public sector with only 12% working in the private sector. (Robinson, 1996: 4)

However, the era of the large-scale, macro-economic policies of development, when the government formed and executed huge national development plans more or less irrespective of the cost is over, and a more prudent approach to labour planning is looking to the private sector to play a larger role, with the threat of compulsory Saudisation targets as an incentive. One of the reasons for their reluctance is the school leaver's lack of occupational skills in general, and English language skills in particular. Employers are reluctant to meet the additional costs of extensive staff training programmes for nationals when expatriates are already qualified and in most cases have lower salary expectations, and so welcomed the development of cost-effective ways of assessing local applicants across a range of abilities, with proficiency in English being an important component in the selection process.

English language skills

The massive influx of immigrant workers had come from many countries, primarily from India and Pakistan, but also Europe and America and South East Asia, to perform all those jobs for which Saudi nationals could not be recruited. The common link between these expatriates was their ability to use English fluently as a working language, either as native speakers or as second or fluent foreign language speakers. Most were unable to speak Arabic and might work in the country for years without any real incentive to learn more than a few words of Arabic. Many of the jobs now being

¹ The Moslem calendar year equivalent to 1994-95 AD

Saudised therefore require a greater or lesser command of English as an international language, to communicate with remaining expatriates and, in an era of increasing globalisation, with the rest of the world.

The kind of English that is needed therefore is a very practical, day-to-day operating language, across the full range of fields of activity: government, banking, manufacturing and the oil industry as well as defence. There is a clear target language use domain (Bachman and Palmer, 1996) and in theory each post can be described in terms of the specific English language requirements to perform that job competently. Whether recruiting to a training programme which is likely to be conducted at least partly in English, or directly into the workforce, employers want an English language test whose results they can interpret with a degree of confidence which state education system and the school leaving certificates do not enjoy.

Not only the syllabi cause employers concern, the general methodology of teaching that favours imitation or rote learning as against experimentation or empiricism is also perceived as a difficulty - mainly because many employees find difficulty in subsequently putting into operation the knowledge that they have acquired at school. It is not that Saudis do not learn English. It is in general taught as a dead language. Saudi students do learn English but in many cases they do not have the strategies to operationalise and use that which they have learnt. In some cases reservations are expressed about examination standards. (Robinson, 1996: 17)

Robinson also makes the point that for private sector employers, "skills are a recruitment and not a training issue. In particular, employers do not see themselves as being responsible for fitting people into the world of work" (page 12). From the employer's point of view, any sustained recruitment programme that did not include at least a screening test for English would increase the need for English language teaching subsequently, and so increase both the cost of Saudisation and the delay in meeting their recruitment targets of nationals and in bringing them on as effective members of the workforce.

There is another, more questionable, motivation for using English language tests in the selection process in such a context; they are a quick and efficient way of 'weeding out' a large proportion of applicants. This 'deselection' is not described in the literature but is an inevitable result of the relative ease and low cost of English language testing compared to the high cost of other formal selection procedures, such as personal interviews and professional skill assessments.

4.2 Five Star test description

The Five Star test was developed specifically to fill this gap for a direct test of English language performance that could be used as part of a larger recruitment and selection process. It is a computer-based test of English language proficiency developed by SDT specifically for use in the Saudi Arabian context, and is reported in Pollard (1994, 1997, 1998a, 1998b, 1999) and Pollard and Underhill (1996). The piloting of the prototype test started in 1993. It has now been used with over 1000 members of the target population - young, male adults seeking to enter the job market - mainly within the parent organisation, but also on a pilot basis with other companies.

"The mission [of the company Saudi Development and Training] was 'Saudization', focusing on the large numbers of Saudi nationals who were in or waiting to enter the workforce. English language was to be an important factor in the services offered and projects tackled. Market research indicated the need for an English language proficiency test for placing people in jobs and vocational training." (Pollard, 1994: 36)

Like many current English language tests, the Five Star test reflects some of the distinguishing features of the communicative approach, but it appears to be the first test that combines them with live interaction in the framework of a computer-based adaptive

test; it employs a lot of authentic and semi-authentic data as input to meaningful tasks, most of which require production or comprehension at discourse level.

Describing the target candidate that emerged from the population profiling exercise, the principal developer of the Five Star Test saw strategic competence as an alternative criterion to native speaker competence:

The SLA [second language acquisition] profile that emerged could be expressed as a continuum from the school-leaver who had never used English outside the Saudi classroom to managers who had travelled and/or lived in EL1 [i.e. English native speaker] environments. Where the latter was true and the EL1 experience was recent to within 2 years, pragmatic and sociolinguistic 'nativeness' might be appropriate. However, there was a much larger group of the population who apparently occupied and dispatched the duties of equivalent positions, but who had either not travelled recently or at all. Since they competently fulfilled the full range of EL [English language] functions in the local environment, 'nativeness' of pragmatic and sociolinguistic forms would seem to be inappropriately prescriptive markers of proficiency. An approach has therefore been adopted whereby strategic competence, defined as coping strategies extending over pragmatic and sociolinguistic competence, has been adopted for completion of the prototype test (Pollard, 1994: 40)

This identification of competence of performance in the local environment as a criterion for judgement is in contrast to the often uncritical use of the native speaker as the yardstick for measurement: "the role of the test is to show how far [the transitional competence of the learner] has moved towards an approximation of a native speaker's system" (Morrow, 1979: 145)

The target candidates live all over Saudi Arabia and are seeking to enter employment or vocational training programmes. They range in proficiency from those who have had exposure to English only through limited formal instruction in government schools to those who may have travelled extensively in Europe or America and have at least a conversational fluency. In the larger metropolitan areas, they will have had an increasing exposure to a variety of regional forms of English in shops, in businesses,

and via the media; access to English language radio, terrestrial television and satellite TV channels is now widespread.

A profile of the target population was drawn up using some 70 questionnaires followed up by structured interviews and a small number of candid recordings (English in use by exemplars of the sample population) and analysis of this data led to a test blueprint including tests of grammar, listening, reading and gapfill skills. It also contained

a test of spoken proficiency, in the form of a structured one-to-one interview... the interview was to be of the traditional structured type - a monolingual test population conspired against group exercises (in-tray) and constraints of time and logistics ruled out task-based lego-construction type tests" (Pollard, 1994: 41).

Group exercises were considered to be inappropriate because participants would naturally use Arabic, their common first language, as a means of communication, and insisting on the use of English would be unnatural. The interviews included some trialling of early task prototypes. The variation in previous exposure to and current fluency in English indicated the need for a general test of proficiency, able to accommodate candidates at all levels of proficiency from beginner to fluent speaker, and an analysis of the questionnaires suggested that the subjects fell into five broad bands of proficiency - hence the name Five Star test (Pollard, 1994: 37).

A tension between the impracticality of conducting conventional oral interviews and other sub-tests with hundreds of learners and questionable validity of indirect methods of assessment was resolved by combining the sub-tests together into a single test event and adopting a computer-based model. The design of the pilot test was originally premised on the testing of six constructs: interaction, listening, speaking, reading, writing and study skills, and all the language skills were to be assessed at this single event, a format suggested in Underhill (1987) and van Lier (1989). However, early tests

showed that the type of test event and the computer-based tasks were not suitable for testing writing, and a separate writing test was developed when the target language use required it. This has subsequently been developed into a fully-functional component of the upgraded version of the test. The writing skill and the original tasks designed to test writing were therefore excluded from this research.

The test takes place between a single candidate and a single interviewer, who combines the roles of assessor (who makes judgements on candidate performance) and interlocutor (someone who engages the candidate in conversation to elicit the language sample on which the judgement can be based - see glossary). Where such oral tests typically take place face-to-face, the Five Star test has the two participants sitting side-by-side at a computer.

The idea of computer-resourcing the test was originally developed with the extended interview component in mind... Once the advantages of computer-resourcing had been established, it wasn't a great step to see how skills such as reading and listening - and, indirectly, writing - could be incorporated and tested in the same process. (Pollard, 1994: 42)

The differing contents of the tasks and the extent and type of interaction they create may lead to a direct collaborative focus on the screen or a shift in posture to a more face-to-face orientation for extended discussion.

The resulting test offers a unique combination in being a) adaptive, b) computer-based but interviewer mediated, and c) heavily dependent on interaction with a live interlocutor. In the pilot form, the test was delivered using Hypercard software on a Macintosh computer. This is a programme that allows 'cards' or 'screens' to be constructed and linked together so that they can be run in sequence, or in one of a number of planned sequences according to the choices made by the operator. In the Five

Star test, the choice is dictated by the interviewer's assessment of the candidate's performance on each task.

Hypercard consists of a number of screens, likened in the literature to cards in a card index. Each card may contain a number of features. For example: sound, graphics and text which can be permanently present, appear and/or disappear after a pre-determined time period, or be triggered manually by clicking on an icon just like the 'pop-up' and 'pull-down' menus of any windows environment. (Pollard, 1994: 43).

The test has been revised and transferred to a CD-ROM for delivery on a different computer platform, but this version has only recently become available (spring 2000).

This research project was based on the pilot version delivered by Hypercard on a Macintosh computer, and the distinctive features that this computer platform allowed are tabulated in Table 3.

Table 3**Distinctive features of Five Star computer platform**

Features	Mode of use (applies in each case some tasks only)	Benefit
Pop up text windows	To give general task instructions, specific question prompts and assessment criteria/correct answers to the interviewer	Reduces memory and processing load on interviewer; standardises administration of task
	To give task instructions to candidate in Arabic	Ensures understanding of task; avoids possibility that candidate has failed task through misunderstanding
	To provide help options on all cards	Ensures focus on completion and assessment of task, not on how to perform it
Timing facility	To provide a pre-determined period of exposure to particular text or graphics; or timing of elapsed response period	Standardises exposure to task stimulus across candidates; allows timed tasks; stopwatch function
Graphics	To present simple 'line-drawing' pictures, graphs, charts	Presentation of visual stimuli is efficient and avoids use of language
Sequence facility	To present simple animated cartoon stories	Allows tasks to be based on sequences of events
Digitised sound	To present instructions in English or Arabic; delivery of aural comprehension tasks	Ensures identical delivery of listening passages; easily repeated; Arabic instructions ensure comprehension of more complex tasks
Progression and scoring by mouse click	To enable the interviewer to record assessment of a task and move to next task by clicking on one of three 'score' buttons	Reduces assessment load on interviewer; computer selects and presents next task according to algorithm
Branching algorithm	To create an adaptive test	Creates more efficient targeting of task difficulty to proficiency level of individual participant
Computer-based recording of assessment	To save to disk an incremental record of interviewer's assessments of candidate performance on each task, calculate the skill profile and generate a report at end of test	Generates instantaneous score profile reporting system

The test is administered to a single participant at a time, and consists of a selection of screens presented on the computer but mediated and scored by the interviewer, who clicks one of three exit buttons at the bottom of each screen. All the computer operations are performed by the interviewer. These screens are described here as distinct tasks, following the discussion of task-based methodology in section 2.14.

The act of scoring each task via mouse-click invokes the computer algorithm which uses the information to select the next task for the test. Typically, a test administration will involve between eight and 15 tasks and last between 10 and 40 minutes. Tests with candidates of higher levels of language proficiency will involve more tasks and take longer than those for more elementary users of English.

The algorithm which determines the sequence of tasks was programmed manually, and all the possible routes (sequences of tasks) are therefore predictable, although there is a very large number of permutations. Part of the algorithm is shown in Appendix III. However, all routes start with the initial task *1-4 Names* and therefore all candidates take this task. It appears to draw its inspiration from a recommendation by Lazaraton who noted the authenticity of the interaction in the introduction sequence used in the OPIs she analyzed, but which is ignored for rating purposes:

A practical suggestion ... would be to use a written agenda or some other written form in the opening segment where introductions occur. Introducing oneself, spelling one's name, and recording it on a form are authentic tasks in which participants routinely engage and which can make the contact less bureaucratic and more personal. (Lazaraton, 1992: 382)

In some cases, the routes allow a topic to be pursued or developed over a series of tasks (Pollard, 1998a). For example, three of the tasks in Table 4: tasks *1-4 Names* (elicitation of names and discussion of family background), and either *3-6 School/study 1* or *4-7*

School/study 2 (discussion of school career) can lead into *11-15 Student reports* and/or *14-19 Reading 4* (listening or reading task about school grades). The system of task numbering is explained below Table 4.

There are in total 73² different tasks, requiring a range of observable language behaviour. These are itemised in Table 4. Some examples of types of language behaviour are answering questions, either spoken by the interviewer or read from the screen; holding an informal conversation on a topic of local relevance; listening to a recorded passage, and answering questions, summarising or selecting specific information; ordering and describing a sequence of events illustrated on the screen; interpreting information presented as a table or graph; filling gaps in passages presented on the screen; and many others. Each task is presented in the context of a topic, and the full range of tasks may therefore be best illustrated in a table which identifies both the language behaviour and the topic that is sampled by each.

It is worth looking at the contents of the test in some detail, in order to appreciate how far it matches the criteria for communicative testing identified in chapters 2 and 3, and where the problem areas remain. Table 4 lists all the operating tasks in the test. Examples of the actual task cards, as they appear on the screen, are given in Appendix II.

² The pilot version of the test actually contains 78 differently-numbered tasks. Of these, one (*X-18 student grades* in table 4) is not linked in to the operating algorithm, so is never used, and there are four sets of identical duplicates, where for programming reasons, the same task is used on two different branching routes and so figures as two different cards. For validation purposes these are treated as a single task.

Table 4 Analysis of Five Star tasks and scoring criteria

Name of task	Topic	Language activity	Scoring criteria
1 4Names	Candidate & family names	Elicit, spell & discuss names	Comprehension/pronunciation: Hesitant, complete, expansive
2 5Base numeracy	Base numeracy	Counting; listen to and identify numbers	Accuracy & fluency : count of correct answers
3 6School/study 1	Candidate's school history	Question & answer; simple questions	Comprehension / fluency: - , partial, complete
4 7School/study 2	Candidate's school history	Question & answer; more complex questions	Comprehension / fluency: - , partial, complete
5 8Basic reading	Basic reading (no single topic)	Read words & sentences	Pronunciation / word recognition: - , acceptable, clear
6 10School/study 3	English learning experience & use	Question & answer; more complex questions	Comprehension & communication:- , partial, clear
7 11Inter numeracy	Intermediate numeracy	Read numbers & simple equations; ,listen to and identify larger numbers	Accuracy: % count of correct answers
8 12Family/recreation	Family & recreation	Question & answer; expand where possible	Interaction: - , partial range, total range
9 13Al Harbis	The two Al Harbi brothers	Listen and re-tell	Comprehension: no duality, duality / no contrast, contrast
10 14Advanced numeracy	Advanced numeracy	Listen to and identify numbers in passage	Accuracy of identification: - , moderate, excellent
11 15Student reports	Student reports / grading	Listen to reports and match against grades	Placement (number correct): 1 of 4 only, 2 of 4, 3 or all 4
12 16Paper clips	Routine office tasks	Listen to instructions; match against priority	Prioritizing: (number correct): 1, 2 or 3, all 4
13 17Reading 3 - Jeddah	City of Jeddah	Reading paragraph aloud	Reading/comprehension: - , partial, fluent/complete
- 18student grades	Table of students' grades	Candidate to talk about table	(no criterion) : - , great difficulty, limited, completes task
<i>nb this task is not included in the computer algorithm, and so does not form part of the test data analysed below</i>			
14 19Reading 4 - grades	Short text on students' test scores	Read aloud and comprehension questions	Reading comprehension: - , adequate, good
15 22Shapes 1	Shapes; slightly differing boxes	Listen and identify graphics of boxes	Accuracy: - , 3 to 5 correct, 6 or 7 correct
16 23Vehicles 1	Features of two new cars	Listen and re-tell	Comprehension: no duality, duality/no contrast, contrast

Table 4 (continued) Analysis of Five Star tasks and scoring criteria

17 24Footballers	Accident between two footballers; timed cartoon story	i) listen to Arabic instructions & explain task in English; ii) watch and narrate incident	Explanation/narration: -, adequate, high accuracy
18 25Ladder	Accident with ladder; timed cartoon story	Order sentences corresponding to graphics	Correct (sentences in order): -, 2 or 3, all 4
19 26Kettle	Accident with child and kettle; timed cartoon story	Order sentences corresponding to graphics	Correct (sentences in order): -, 2 or 3, all 4
20 27Writing*	Simple sentence	Copy sentence on sheet of paper	(writing) : illegible, printed, cursive
21 28Signs	Road signs; timed graphic display	Identify correct text for road sign from 6 time-displayed choices	Text identification: -, average, maximum
22 29Fridge	Instructions for use of fridge; recorded passage and graphics	Explain instructions in own words, from recorded passage and graphic illustrations	Information: -, adequate, high accuracy
23 30Reading 2	Short text on a university student	Read aloud and comprehension questions	Reading comprehension: -, adequate, complete
24 31Writing*	Simple sentence	Copy sentence on sheet of paper	(writing) : illegible, printed, cursive
25 33Traffic lights	Traffic signal instructions	Say what driver should do at different signals	(no explicit criterion) : -, adequate, fluent
26 34Traffic lights 2	Traffic signal advice	Read text and complete gapped words	Completion: -, hesitant, fluent
27 36Signs 2	Airport sign; timed graphic display	Identify correct text for airport sign from 6 time-displayed choices	Text identification: -, average, maximum
28 47Signs 3	Warning message on bottle: timed graphic display	Identify correct text for warning from 6 time-displayed choices	Text identification: -, average, maximum
29 50Road signs	Road sign instructions	Say what driver should do at different signs	(no explicit criterion) : -, adequate, fluent
30 51Road signs 2	Traffic sign advice	Read text and complete gapped words	Completion: -, hesitant, fluent
31 53Training center	Training centre courses	Listen and re-tell	Comprehension/fluency: -, partial, complete
32 54Population	Rate of population growth	Listen and re-tell	Comprehension/fluency: -, partial, complete

Table 4 (continued)

Analysis of Five Star tasks and scoring criteria

33 55Kuwait City	Visit of George Bush to Kuwait	Listen and re-tell	Comprehension/narration: - , partial, complete/fluent
34 56Nagorno Karabakh	Resolution of border dispute	Listen and re-tell	Comprehension/narration: - , partial, complete/fluent
35 57Making tea	Making English tea	Use pictures to explain sequence of instructions	Communication/lexis: - , clear/limited range, clear/full range
36 58Speculation 1	Family size	Discuss pros and cons; respond to challenges	Lexical range/fluency: - , adequate, expansive
37 59Puncture repair	Replacing punctured tyre	Use pictures to explain sequence of instructions	Communication/lexis: - , clear/limited range, clear/full range
38 60Singapore	Report of boat capsized	Listen and re-tell	Comprehension/fluency: - , adequate, complete
39 61Speculation 2	Personal education/ qualifications, aspirations and hopes	Discuss personal ambitions; respond to challenges	Lexical range/fluency: - , adequate, expansive
40 62Speculation 3	Car ownership, ideal car, new vs used car	Discuss car ownership; respond to challenges	Lexical range/fluency: - , adequate, expansive
41 63Road accidents	Causes of traffic accidents	Read text and complete gapped words	Completion: - , hesitant, fluent
42 65Regional affairs	Arab world affairs	Describe one major recent event in the region	Lexical range: - , limited, wide
43 68Newspaper 1	Article about Azerbaijan	Read article silently, answer timed questions	Time taken for 3 questions: > 71 secs, 41 - 70 secs, < 40 secs
44 69Newspaper 2	Article about guerillas in Peru	Read article silently, answer timed questions	(Number of correct answers) : (1-2), 3, 4
45 71Instructions	Word-to-picture matching ex	Translate & explain instructions	Detail accuracy : - , most, full
46 72Lebanon	Newspaper gagging report	Listen and re-tell	Comprehension/narration: - , partial, complete/fluent
47 73Instructions 2	Word-to-picture matching ex	Translate & explain instructions	Detail accuracy : - , most, full
48 74Lille	Industrial decline in Lille	Listen and re-tell	Comprehension/narration: - , partial, complete/fluent
49 75Saudia timetable	Airline timetable	Match explanatory labels (time-limited display) to timetable	Correct (matches) : - , 6 to 8, 9 or 10

Table 4 (continued)

Analysis of Five Star tasks and scoring criteria

50 76Weather charts	Temperature & rainfall charts	Read 5 short texts (time-limited display) and fill info/word gaps from charts	Correct (answers) ; - , 5 to 7, 8 or 9
51 88Riyadh weather	Temperature & rainfall charts	Match explanatory labels (time-limited display) to charts	Correct (matches) : - , 5 to 7, 8 or 9
52 89Climatic change	Climate change	Read jumbled sentences (time-limited display) and re-order	Accuracy: over 120 secs, 4 or 5 correct, 100%
53 91Child death	Tables of child mortality and disease linked to poor food	Read 2 short texts (time-limited display) and fill 5 info/word gaps from charts	(correct answers) ; - , 3 or 4, all 5
54 92Travel	Travel; comparison of 2 countries from personal experience	Discuss countries and fill table with ratings for VFM, climate, scenery, services	Participation: - , limited, balanced
55 94Heathrow	Heathrow airport text	Read aloud and fill gaps (C-test)	(Correct gapfills): - , 6 to 8 correct, 9 or 10 correct
56 97Tim Severin	Tim Severin's journey across Pacific	Listen and re-tell	Comprehension / fluency: - , completes with assistance, complete & fluent
57 98Free money	Currency exchange at airports	Read 6 jumbled sentences (time-limited display) and re-order	(Correct order) : less than 4, 4 or 5, 100%
58 100United Nations	UN report on manufacturing in Indonesia	Listen and re-tell	Comprehension/ vocabulary range: - , adequate, total/ extensive
59 101US Hitech	Easing of US hi-tech export restrictions	Listen and re-tell	Comprehension/ vocabulary range: - , adequate, total/ extensive
60 102UNIDO	UNIDO text	Read aloud and fill gaps (C-test)	(Correct gapfills): - , 8 to 9 correct, all 10 correct
61 103Company priorities	List of general company priorities	Discuss objectives in context of personal experience; match actions (time-limited display) to objectives; respond to challenges	Comprehension/ interaction : - , adequate, complete / wide-ranging

Table 4 (continued) Analysis of Five Star tasks and scoring criteria

62 104Karoshi	Death from overwork	Read 5 jumbled sentences (time-limited display) and re-order	Accuracy: > 180 secs, 2 or 3 correct, 100%
63 105Karoshi 2	Article about overwork	Read article silently, answer timed questions	Time taken for 3 questions: -, 41 - 70 secs, < 40 secs
64 107Prices	Table of changes in consumer/producer prices in different countries	Read 5 time limited sentences about data from tables and fill gaps from list of 'change' words	(correct gapfills): -, 4 or 5 correct, 6 or 7 correct
65 108Production	Table of changes in production and sales in different countries	Read 6 time limited sentences and fill gaps with data from tables	Accuracy(correct gapfills) : -, 4 or 5, 6 or 7
66 109Porsche	Porsche policy and prices	Read text aloud and fill gaps (C-test)	Accuracy (correct gapfills) : -, 7 or 8, 9 or 10
67 110Book review	Review of book on Saudi finance markets	Listen and re-tell	Comprehension/fluency : -, adequate, complete/high
68 112Bosnia	Report on Civil war in Bosnia	Read text aloud and fill gaps (C-test)	Correct (gapfills) : -, 7 or 8, 9 or 10
69 113Conservation	Strategies for saving the planet	Read list of strategies; read two short texts (time-limited display) and explain link to strategies; be able to expand	Comprehension/interaction : -, adequate, complete / wide-ranging
70 114SA railway	Lecture on advantages & disadvantages of national railway system for Saudi Arabia	Read headings (time-limited display) for lecture notes, expand on relevance to lecture topic	Comprehension/fluency : -, adequate, complete/high
71 115Honey bee	Life cycle of honey bee	Read 6 jumbled sentences (time-limited display) and re-order	Accuracy: > 300 secs, 2 to 4 correct, 100%
72 120Conservation	Use of fossil fuels and global warming	Read text aloud and fill gaps (C-test)	Accuracy (correct gapfills): -, 7 or 8, 9 or 10
73 123The computer	Memo about computer repair	Read text aloud and fill gaps (C-test)	Accuracy (correct gapfills): -, 7 or 8, 9 or 10

Table 4 contains four columns. The first column, *Name of task*, gives the title that appears on the screen for each task card, preceded by two numbers. The first number, which runs sequentially from 1 to 73, is the reference number of that task for the purpose of this data analysis. The second number is the card reference number as it appears on the computer screen, and runs from 4 to 123 with gaps for duplicates, scoring cards, redundant and exit cards. Thus, task *1-4 Names* is the first task in the sequence but is numbered card 4 in the pilot test. Task *X-18 Student grades* is not linked in to the test operating algorithm, so was included in the panel scrutiny but not in the test data analysis.

The second column, *Topic*, identifies the topic of the activity. In the great majority of cases, this topic extends beyond the simple sentence or utterance level, i.e. whether in speech or in writing the task provides the opportunity for extended discourse.

The third column, *Language activity*, gives a simple descriptive label of what the candidate has to do in English in order to complete the task. Thus, listening to instructions in Arabic solely in order to ensure comprehension of the task will not be included here, but where a task requires listening to a digitised passage in English, then 'listening' will be identified as one of the activities. This column is based on a superficial analysis of what candidate has to do. It does not pretend to identify which language skill or skills are needed, which is the focus of one part of the expert panel analysis reported on in chapters 5 and 6, but clearly it would be surprising if there were not some common ground between them.

The final column, *Scoring criteria*, gives the marking criterion for each task and the wording on the exit labels. Thus, for task *1-4 Names*, the overall criterion is *Comprehension/pronunciation* and the three exits are labeled *hesitant*, *complete*, *expansive*. As mentioned above, the interviewer scores performance on each task by clicking one of three buttons. In most cases, there is an overall criterion for assessing performance on that task against, for example, reading comprehension, accuracy, or number of sentences placed in the correct order. In a few cases, there is no explicit criterion, but it can be inferred from the nature of the task and the explicit exit labels.

Again, in most cases, the three exits also have distinct labels, to indicate the level of performance needed on the task criterion to achieve that exit. In some cases, there is no explicit label for the lowest of the three exits, which by default becomes 'fails to achieve the performance level required for the second (middle) exit'. Such a case is indicated in the table as -, *partial*, *complete* where the second and third exits are labeled *partial* and *complete* but the first exit has no label.

Scoring system

As each task in the Five Star test is completed, the interviewer makes a judgement on what is effectively a three-point scale and clicks one of the three score buttons at the bottom of the screen. The three potential scorers 'can be characterised broadly as 'non-performance', 'partial performance' and 'complete performance' (Pollard, in progress) and to help the interviewer to make this judgement, pop-up descriptors specific to each scale point are available for most tasks. The single button-click score for each task reduces the cognitive load on the interviewer which in a typical oral test can raise the

affective barrier between interviewer and candidate and render the event less like ordinary conversation. Both the assessor and candidate should be able to behave more naturally if the former is relieved of the burden of cumulative assessment in this way (Pollard 1998a).

The task by task incremental scoring is also more likely to result in a reliable score than a global assessment made retrospectively at the end of a test, and the interviewer does not need to carry over his or her judgements from one task to the next and continually refine them; this is done by the computer. The physical act of scoring is to some extent masked by the fact that the single button-click also causes the next task to appear on the screen.

For each task, the computer program contains a pre-determined allocation of the score across one or more of the skills, in different proportions. For example, the allocation of the score for a task involving a set of interview questions might be 50% to interaction, 30% to listening and 20% to speaking (example from Pollard 1994:52). The other three skills used are reading, study skills and writing, but only four of these were operationalised for the expert panel analyses (section 5.1.2).

Rather than being combined into a single overall assessment, the cumulative scores for each skill are stored in a database as the test proceeds and the final score in each of the six skill areas is reported separately through a profile that resembles a bar graph, with the six skills along the bottom axis and the rating scale from 1 to 5 (hence 'Five Star') along the vertical axis. The profile for any candidate can then in principle be matched

against a similar English language profile drawn up as a result of a needs analysis for a particular job or traineeship that is being recruited for.

Having made these proportional allocations, the test designers then constructed the algorithm so that the task sequences on all the possible routes provided a sufficient sample of each skill, and, as the number of tasks and balance of skills vary from route to route, the programme calculates the final score profiles out of the maximum possible scores on that route. "At the end of the test, when a candidate has worked his particular route, all the scores derived for, say, interaction will be summed and measured against the maximum score that would have been possible for that skill *on the tasks completed*." (Pollard 1994:52).

The allocation of skills tested to each task, and of proportions of the total score across those skills, was necessarily based on intuition at the pilot stage. Part of the purpose of the critical review of the test carried out by the expert panel was to establish the validity of these skill constructs and score allocations as realised by the test tasks, and this is reported on in chapters 5 and 6.

4.3 Summary of key communicative features

Table 5 is a summary of the key features of the Five Star test against the features of the communicative approach identified in chapter two.

Table 5

Summary of key communicative features

Communicative feature	Implementation
Communicative	<ul style="list-style-type: none"> • Almost all the tasks require the processing or expression of real meanings, i.e. statements or propositions relating to real world contexts shared by the participants (candidate and interviewer)
Task based	<ul style="list-style-type: none"> • The 'unit of testing' is the task, each of which may take several minutes, rather than a single test item or question. Each task score is therefore based on extended discourse. • The 'side-by-side' positioning of interviewer and candidate serves to emphasise their joint engagement with a common task, rather than the usual confrontational position of a typical interview.
Integrated skills	<ul style="list-style-type: none"> • Tasks mostly call on complex combinations of language skills rather than a single skill, and this is reflected in the scoring system
Authentic materials	<ul style="list-style-type: none"> • The topic and text (audio or script) of most tasks is taken from the real world context familiar to most candidates; the cultural homogeneity of the target profile makes this much easier to do. • Culturally external data is mostly in the style of news reports.
Emphasis on oral interaction as a reflection of strategic competence	<ul style="list-style-type: none"> • Virtually all the tasks require direct oral communication between candidate and interlocutor, and in principle either party may extend and develop this interaction (it is not entirely predictable). • This negotiation enables the display of strategic competence above the micro-linguistic level of performance • The interaction may also extend over the sequence of tasks in a test based on thematic linkage of tasks or recurrence of topics
Individualised and learner-centred	<ul style="list-style-type: none"> • The one-to-one format and the open-endedness of the interaction make possible relatively life-like conversations between interviewer and candidate. • Some tasks specifically call for expression and justification of personal opinions and truths. <ul style="list-style-type: none"> ▪ Being an adaptive test, a higher proportion of the tasks presented should be at or about the language proficiency level of each individual candidate

Constraints

Topicality The corollary of many of these communicative features is that the test is specific in culture, space and time to the context for which it was conceived: the Arab world, specifically Saudi Arabia, in the early-mid 1990s. If tasks are to be relevant to contemporary issues immediately outside the test event, then maintaining geographical and temporal relevance will be a continuing challenge; merely 'doing nothing' for a period will undermine construct validity based on these communicative principles. What may be judged valid one year will not be so five years later. Examples of tasks (see Table 4) that already appear dated are

33-55 Kuwait City: Visit of George Bush to Kuwait ,

44-69 Newspaper 2 : Article about guerillas in Peru

68-112 Bosnia : Report on Civil war in Bosnia.

Rapid social and cultural changes might affect the prominence given to the role of the motor car or the complete lack of mention of religion or the place of women in local society. Although linguistically appropriate for a wider regional market, the Saudi focus of many tasks means they would lack the immediate relevance, and hence the construct validity, for candidates even in neighbouring countries with similar demographic issues such as Kuwait or the Emirates.

The problem of topicality in communicative tests is taken up again in chapter seven. Part of the purpose of this research is to identify a validation process that can accommodate the continuing updating needed to retain communicative relevance and construct validity.

Physical constraints

A different group of constraints are the results of specific physical features of the Five Star test.

- a) The situational context of the test event precludes other permutations of interaction and skill; the interaction is at all times between just two people, in the fixed roles of candidate and interviewer. Other tests allow a range of patterns of interaction between different people and roles.
- b) The test is therefore labour-intensive, an interviewer being required for the duration of each test event on a one-to-one ratio. Other tests use a more economical staffing ration particularly when listening and reading are being tested in ways which do not require the same face-to-face interaction as speaking.
- c) Being computer-based, authentic writing tasks are inappropriate and would greatly increase the length and hence the cost of the test. Other tests that include such tasks have separate writing papers which may last an hour or more. As will be seen in chapter five, writing was eventually excluded from the list of skills evaluated in the panel exercise.

- d) The interviewer combines the role of interlocutor (someone whose job is to engage the candidate in conversation, and to encourage the candidate to make the most of his/her oral fluency) with the role of assessor (whose job is to make a judgement of the candidate's performance on the given criteria as objectively as possible). To carry out these roles simultaneously is demanding, some would say impossible; some other tests separate them.

To see how other tests with claims to communicative construct validity deal with these issues, it is instructive to look at a comparison with other current tests of speaking.

4.4 Comparison with other current tests of speaking

To place the Five Star test in the context of current operational testing, Table 6 summarises features of five commercially available language tests (columns 1 - 5) which include the elicitation of spoken language, compared against the Five Star test (column 6).

It shows how the Five Star test shares many of the distinctive features of oral tests that have evolved from the communicative approach, but also how it is distinctively different in other respects.

Table 6 Other current oral test formats

Column	1	2	3	4	5	6
Row 1 Name	Spoken English (Grade Exams)	IELTS Speaking	CCSE Speaking	Cambridge CAE Paper 5 Speaking	Test of Spoken English (TSE)	Five Star
Row 2 Source	Trinity College London, UK	British Council/ Cambridge Exams Syndicate/ IDP	Cambridge Exams Syndicate, UK	Cambridge Exams Syndicate	Educational Testing Service, USA	SDT, Saudi Ar
Row 3 Live or recorded? Adaptive?	Live; test format controlled but some scope to vary interaction	live; test format controlled but some scope to vary interaction	Live; test format controlled but some scope to vary interaction	Live; test format controlled but some scope to vary interaction	scripted and recorded : 'semi-direct'	live; test is ada and scope to v interaction
Row 4 How many people are involved? Who?	2: candidate, interviewer	2: candidate, interviewer	3 or 5: one or two candidates, assessor, interlocutor, usher	4: two candidates, two examiners	1: candidate speaks to tape which is marked by assessor subsequently	2: candidate, interviewer

Table 6 (continued)

Other current oral test formats

Column Name	1	2	3	4	5	6
	Spoken English (Grade Exams)	IELTS Speaking	CCSE Speaking	Cambridge CAE Paper 5 Speaking	Test of Spoken English (TSE)	Five Star test
Row 5 What test techniques/tasks are used to elicit spoken language?	<ul style="list-style-type: none"> • Conversation • Presentation of prepared topic • Discussion of listed subject areas • Presentation of 1 or 2 texts • Discussion of listening comprehension 	<ul style="list-style-type: none"> • Introduction • Extended discourse (explanation, description, narration) • Elicitation • Speculation and attitudes • Conclusion 	<ul style="list-style-type: none"> • Discussion based on task • Interaction between 2 candidates • Reporting back • Discussion of other related areas 	<ul style="list-style-type: none"> • Introduce and respond to questions about self • Describe & comment on visual prompt • Problem-solving task • Report back; wider discussion 	<ul style="list-style-type: none"> • 'warm up' questions • answer questions about map • tell picture story • discuss topics of general interest • describe info in graph • present info from schedule 	<ul style="list-style-type: none"> • adaptive sequence of tasks stored on computer • tasks test different skills singly and in combination but most tasks require spoken output (see in chapter 2)
Row 6 Duration of test	5-25 minutes	10-15 minutes	10 mins preparation 15 mins test	15 minutes	12 items of 30-90 secs, total 20 mins	8-15 tasks, total 10-40 mins

Each of the six rows in Table 6 is now described in detail.

The first two rows give the name of the test and the examination board. The Five Star test is described in column 6. The other five tests are:

Column 1: The Examinations in Spoken English Grade Exams from Trinity College
London

Column 2: The Speaking module of the International English Language Testing System,
jointly assessed and run by the University of Cambridge Local
Examination Syndicate (UCLES), the British Council and the
International Development Program of Education Australia (IDPEA).

Column 3: The speaking component of the Certificates in Communicative Skills in
English, from UCLES (formerly called 'oral interaction' - UCLES, 1995)

Column 4 : The Speaking Test from the Certificate in Advanced English, from UCLES

Column 5: The Test of Spoken English (TSE), from Educational Testing Service,
Princeton, New Jersey, USA

The Trinity College exams (column 1) test only spoken English. The TSE (column 5) tests only spoken English, but is indirectly linked to the Test of Written English (TWE) and Test of English as a Foreign Language (TOEFL) which test other aspects. The other tests considered here (columns 2, 3 and 4) are in effect the speaking sub-tests of a battery of four or five tests. Five Star is unique in claiming to tap integrated language skills in a single test event.

These exams have been selected for the comparison on the basis that they are

- a) available internationally on a scheduled or 'on demand' basis;
- b) they all lay claim to test communicative ability in spoken English; and

- c) they are commercially operational, in the sense that they are routinely administered on a relatively large scale, with several thousands or tens of thousands of candidates each year.

There many other language examinations which could have been included in this table, but it would be impractical to attempt to be comprehensive. The fact that three of the five comparators are from UCLES is an indication of the dominance of the Cambridge Examinations Board in this field. The only other board that exceeds their annual statistics for candidates is the Educational Testing Service, whose Test of English as Foreign Language (TOEFL) is the biggest and best known of all English language exams, claiming approximately 700,000 candidates per year world-wide (Netten, 2000b) but which contains no spoken component. It is in its current form the archetypal multiple-choice test and contains no speaking component of any kind, but the development plans described in more detail in the following section suggest that it too is being influenced by the communicative revolution.

In other words, they are looking at incorporating key communicative features of the Five Star and other tests listed below. The possible evolution of the archetypal multiple-choice test in this direction would seem to vindicate the principled adoption of communicative methodology in even large-scale language testing. Although the theoretical requirement for tasks that test integrated skills has been a cardinal principle of the communicative approach for many years (section 2.9.2), it seems to be the transfer to a computer-based platform that is actually making possible the development of such tasks, both for the Five Star test and TOEFL 2000 (Netten, 2000b).

The third row of Table 6 shows whether the speaking test is live (a direct test) such as Five Star or recorded (semi-direct), and to what extent it is adaptive. As noted in 3.4 above, a live semi-structured oral interview can be considered adaptive in the sense that the interviewer will naturally vary the difficulty of at least part of the tasks in order to focus on the approximate level of the candidate's proficiency; it is inefficient to spend time on tasks that are clearly too easy or too difficult. In this sense, all the direct oral tests in columns 1 - 4 are to some extent adaptive.

However, there is a distinction between tests that a candidate enters at a particular level, and the assessment only determines whether s/he has met that level, leading to a pass/fail outcome (columns 1, 3 and 4); and tests such as Five Star that assess performance against a broad scale, with the outcome reported as a score or band on that scale, and without a pass/fail judgement being made (columns 2, 5 and 6). There is obviously more scope for adaptive interviewing where there is a wide range of possible outcomes rather than a single pass/fail criterion.

The fourth row of Table 6 shows how many people are involved in the test. The terminology follows the convention that someone who only interacts with the candidate and does not make an assessment is called an 'interlocutor', someone who assesses but does not interact is an 'assessor', and someone who plays both roles is an 'interviewer'. The Five Star pattern of a single interviewer and a single candidate is found also in columns 1 and 2, but other exams separate the role of assessor and interlocutor (column 3) or have two interviewers and two candidates (column 4). Research underlying the

varied patterns of interaction (Ffrench, 2000) in the Cambridge 'main suite' exams (UCLES, current) is described in section 3.3.2 above.

The fifth row identifies the type of techniques or tasks used to elicit the speech sample. It is beyond the scope of this survey to attempt a comparison of the topics, although one of the appendices of the critical review (Underhill, 1997) describes a content comparison of the Five Star with Trinity College, IELTS and Cambridge main suite (columns 1, 2 and 4) in more detail (Appendix IV).

What is evident from row five of the table is that all the oral tests considered here contain a sequence of tasks or activities that present and elicit language in different ways. None of these tests relies on a single method of elicitation. The arguments usually put forward for building in such a variety of speaking activities include authenticity (it is more life-like); fairness (different candidates will be better at different types of tasks); balance (more communicative tasks are mixed with more mechanical ones, more subjective with more objective); and flexibility (there is more scope for adaptivity and adaptability to circumstances, environment and resources) (Underhill, 1987: 38) However, the Five Star test is unique in having so many distinct tasks, between eight and 15 in a test event, compared to the usual four or five in the other tests.

The Five Star is also unique in that all the prompts, whether text, visual, or audio-recorded are presented by the computer, rather than by the interviewer, and this was seen at the outset as a real advantage:

[Conventional] oral interviews are fraught with difficulties. Maintaining a scoring criteria involves either print-outs of band descriptions, 'tick-the-box' marking matrices, or phenomenal powers of memory and concentration. The former are a source of unwanted distraction for tester and learner, and are subject to interpretation; the latter are extremely scarce and susceptible to fatigue. If discourse prompts such as pictures are used, these have to be organised and accessed efficiently; for even the most adept interviewer this will interrupt both concentration and conversation. These are amongst factors likely to contribute to poor inter- and intra-rater reliability and other sources of error measurement. (Pollard, 1994: 42)

The sixth row of Table 6 indicates the duration of the test. Rather than fixed duration, ranges are reported for the Trinity College, IELTS and Five Star tests (columns 1, 2 and 6), because these are the live tests that are applicable to any level of proficiency, and the duration reflects directly the proficiency level of the candidates. Test events for more proficient candidates take longer than test events for less proficient ones.

4.5 Comparison with other computer-based tests

Early use of computers in language testing was largely restricted to the analysis of test results, but greater availability of computer hardware and ease of use of programming software has led to their widespread use for test delivery. The reduction in cost of personal computers in particular has made possible the development of programme-specific tests in schools and language teaching programmes which are not publicised or made commercially available, and about which it is therefore difficult to get information. The examples of computer-based tests described here are therefore all commercially-produced tests.

On a terminological note, Alderson defines CBELT (computer-based English language tests) as "tests which are delivered and scored by computer" (Alderson, 1990: 22).

whereas tasks in the Five Star test are delivered by computer but mediated and scored by the interviewer The first five examples below are all scored by computer.

- i) ToPE: The Test of Proficiency in English (Hill, 1990, 1995) is a computer-based adaptive test that uses a gapfill technique. ToPE has about 500 items in total, each item being a reading text of 150-250 words, from which words are deleted at fixed intervals, and it is the candidate's task to supply the deleted words via the keyboard. This is the only type of task. Correct answers only are accepted, disallowing what might be considered acceptable alternatives; this is a contentious issue in cloze and gapfill language tests, but in the context of a computer-marked test, single-correct answer is the only feasible choice. Certain words, such as proper nouns and numbers, are excluded from the deletion process. This interval varies from text to text, which were selected from a wide variety of sources such as newspapers, reference books, short stories, and teaching texts.

Each text in ToPE is treated as a single 'task' with a pre-determined level of difficulty, and an underlying pre-programmed algorithm determines the route through the test. Results are reported on a nine-band scale chosen to fit with the English Speaking Union's Framework (see Table 11), as used in the Five Star panel exercise. The test event finishes typically after three or four texts have been completed, and takes from 40 minutes to one hour, although no specific time limit is set.

- ii) CommuniCAT is an adaptive test published by UCLES (University of Cambridge Local Examination Syndicate). A large item bank contains test items of different kinds, such as multiple choice listening and reading, cloze (gapfill) and sentence transformation, and employs colour, graphics and digitised sound files.

All items are dichotomously scored (right or wrong). The computer immediately assesses a candidate's response and chooses the next item on the basis of random selection from items of the appropriate level. It is claimed that 'as few as 18-20 items can accurately define a candidate's level of ability' (Williams, 2000). The test event is terminated when the estimate of error reaches certain pre-defined confidence limits. There is currently no assessment of speaking skills and limited assessment of writing skills, although plans are in hand to address these limitations in future.

Versions are also available to test French, German and Spanish and offers multilingual instructions to users. A report is issued as soon as the test is completed, which relates the candidate's level of ability to the five 'main suite' proficiency levels (UCLES EFL website and 'CommuniCAT' handbook, current) and can also include 'can-do' statements developed with ALTE (Association of Language Testers in Europe) (ALTE website)

- iii) Computer-based TOEFL: The Test of English as Foreign Language is the most commonly used criterion for evidence of English language proficiency for non-native speaker applicants to colleges and universities in the United States, and

claims 11 million candidates since it was introduced in the 1960s. A computer-based version was introduced in 1998 and is now taken 'on demand' by around 300,000 of the 700,000 candidates annually. Advantages of the computer-based test cited by TOEFL include testing on demand by prior appointment, whereas the paper-based test is only available on fixed dates, and immediate viewing of scores, except for the extended writing task (Netten, 2000b).

However, plans to move all TOEFL testing to the computer-based version and to discontinue the pen-and-paper test have had to be suspended or at least postponed, due to difficulties providing and supporting computer testing facilities in a number of developing countries where it can only be taken on monthly scheduled dates.

The test is a multiple-choice format containing two major components, reading and listening. The paper-based version is entirely linear, but sections of the computer-based version are adaptive; the reading section, however, is not adaptive, as it contains sequences of questions based on the same text, which could not be transferred to an adaptive format, and also because this would violate the assumption of independence between items that underlies the IRT model used. This combination of linear and adaptive bases in one test is unusual, perhaps unique. TOEFL have published numerous research reports and other literature about the computer-based test and its equivalence with the pen-and-paper test (e.g. ETS 1998a, 1998b, 1999, TOEFL website) but it does not say how the models combine, in other words, how scores derived from the linear sub-test are combined with scores derived from the adaptive sub-tests.

There is no performance test of speaking, but the TOEFL can be combined with the semi-direct TSE (Test of Spoken English) also published by the Princeton-based Educational Testing Service. The computer-based TOEFL contains in addition a short essay writing task. Future plans include the development of tasks that are more communicative and that test integrated skills. (Netten, 2000a; Netten, 2000b; TOEFL website). Changes already announced have heralded a move away from discrete items and towards a more integrative approach, such as

... eliminating single-statement ...items, expanding the number of academic lectures and longer dialogs, and embedding vocabulary in reading comprehension passages.... The TOEFL 2000 project is a broad effort under which language testing at ETS will evolve into the twenty-first century. The impetus for TOEFL 2000 came from the various constituencies, including TOEFL committees and score users. These groups have called for a new TOEFL test that is (1) more reflective of models of communicative competence; (2) includes more constructed-response items and direct measures of writing and speaking; (3) includes test tasks integrated across modalities ... (ETS, 1998a: 2)

- iv) PhonePass is a test of conversational oral proficiency in English which is administered by computer over the telephone, based on algorithms for the computer analysis of speech. Candidates do not see or physically operate the computer, but they hear a series of supposedly interactive tasks, such as to repeat a sentence or answer a question. The test lasts 10 minutes and has five components: read aloud, repeat sentence, opposite word, short answer and open response; the first four parts are scored automatically by machine and in some cases there is more than one possible correct answer. The fifth part collects two 30-second samples of speech which can subsequently be reviewed by score users (Ordinate 1998; Ordinate website). The test is not adaptive.

A validation report claims concurrent validity correlations of .75 against scores of 392 candidates on the TOEFL test (see above) and between .52 and .77 against five different speaking tests administered to samples ranging from 51 to 171 non-native speakers (Ordinate, 1999). None of these coefficients is so high as to provide convincing criterion validity, and it is difficult to justify the publisher's claim that "these results are generally consistent with the hypothesis that PhonePass scores roughly correspond to oral proficiency ratings" (Ordinate, 1999: 4) even with the word 'roughly' as a qualification.

Because it can be taken by appointment from anywhere in the world and needs only access to a telephone rather than a computer system, it is cheap and easy to take and convenient to administer. It is a semi-direct test which claims to be interactive and does indeed elicit spoken samples of language but suffers from the drawbacks described in 7.2.1 below: no real two-way communication, a restricted range of speech functions and a lack of face validity. It is not recommended to measure or differentiate candidates by advanced skills (Ordinate, 1998) suggesting that the speech analysis technology is still at a stage where it can convincingly cope with speech at the word and sentence level, but not at the discourse level or over the full range of language functions that direct tests can potentially measure.

- v) CATS (Computer Assisted Training System) Corporate English Test is a multiple-choice test for learners of business English. It is an integral part of a larger training package, and the test is taken over a number of test sessions, after each of which a candidate's results are analysed and recommendations for

study using the Corporate English Training Materials are made - the package includes over 500 computer-based 'training modules'. A diagnostic report from each test identifies questions answered incorrectly, together with the correct answer and the student's answer. The selection of recommended course materials is personalised, based on the diagnostic test results, but the test itself does not appear to be adaptive nor are there any performance skills tested or task types other than multiple choice (CATS website).

Of the five computer-based tests mentioned here, only one, PhonePass, elicits actual spoken language from the candidate, and that is in a semi-direct format that is analysed by machine and is not adaptive. Two institutional-specific examples of computer-based semi-direct oral tests which are recorded and scored later by human assessors, are COPI (Computerised Oral Proficiency Interview) in Spanish, Chinese and Arabic under development at the Center For Applied Linguistics in Washington D.C. (Malabonga, 1998) and CCPE (Computer-Assisted Communications Placement Exam) in English containing self-introduction, picture description and video-lecture tasks at Nassau Community College (Gulinello and Durso, 1998). Although some are technically more sophisticated than the Five Star test in terms of the algorithm underlying task selection and sequencing, none of these tests displays the combination of communicative features that the Five Star test does or has any realistic claim to test integrated language skills.

The crucial feature of the Five Star that is absent from these other tests is the human interlocutor. Although the elicitation of speech is not itself ruled out in such tests through speech analysis or semi-direct recorded oral test activities, the term 'computer-based' seems to be interpreted as dispensing with the need for a live interviewer.

Most recently, the spread of the internet has led to the development of web-based testing, and an extensive and up-to-date list of links to web-based tests can be found at the Resources in Language Testing website and there is a comparative analysis of some web-based testing systems on the North Carolina State University (NCSU) website (Gibson et al., 1996).

4.6 Summary

The Five Star test was developed in and is tied to a very specific geographical and cultural context. While this may be a positive feature in terms of communicative testing criteria (section 2.10), it means that any validity that is established for the test and the use of its scores would have strictly limited transferability to other test populations in other contexts. The use of spoken and written Arabic language in many tasks would alone narrow the geographical range of possible applications.

The Five Star test incorporates several features of communicative testing that were identified in chapters 2 and 3, but does not include the paired or group testing of participants adopted by UCLES. Many of the techniques used to elicit spoken language are also used in other current tests of spoken language. However, it is unique among these oral tests in setting out to test integrated skills, in being explicitly adaptive so that different candidates may be presented with different tasks, and in using a computer system to deliver these tasks.

Compared against other computer-based tests, the Five Star test has less in common. It is the only one with direct testing of spoken language, which is made possible by the presence of a human interlocutor. The absence of this interlocutor in other computer-based tests may be precisely because computer delivery is seen as an opportunity to introduce more sophisticated test tasks than conventional pen-and-paper activities without the expense of a human administrator.

The uniqueness of the Five Star test in these comparisons lies not in any one aspect but in the combination of a number of attributes, such as live oral interaction and other features of the communicative approach, integrated tasks in a strong cultural context and an adaptive algorithm delivered by a computer. In this respect the present research attempts to validate the Five Star test not as the sole objective but as an exemplar of a larger class of tests. made possible by recent developments in methodology, test theory and the availability of computer systems, which share some but not necessarily all of these attributes.

Among the constraints implied by the attributes are likely to be a modest scale and modest resources for validation and evaluation, entailed by the specificity of local context, and the need for continuous validation, entailed by the need to constantly update the test content to ensure topicality and authenticity. The aim of this research is to develop such a model for continuous validation, and the next chapter looks at two different possible source of data for it.

Chapter 5 Data collection and research methodology

- 5.0 Introduction
- 5.1 Data set 1: Expert panel
 - 5.1.1 Expert panel: description of methodology
 - 5.1.2 Expert panel: data collection
 - 5.1.3 Content validation against external criteria tests and development of tester orientation package
- 5.2 Data set 2: Item response theory
 - 5.2.1 Item response theory: description of methodology
 - 5.2.2 Item response theory: data collection
- 5.3 How has the methodology developed over the project?
- 5.4 Alternative research methods
- 5.5 Summary

5.0 Introduction

This chapter describes the methods used for data collection and the principles underlying the methodology for the two large data sets used in this research. Chapter six then analyses and discusses these data.

The first dataset, described in section 5.1, consists of scrutiny by an expert panel of the pilot version of the test. Although later stages of the panel exercise included viewing video records of authentic test events, the main part of the panel's work was based entirely on the test instrument itself done independently of any actual test results. An

additional activity carried out within the framework of the panel exercise was a content comparison against some external tests, and this is described in section 5.1.3 and given in full in Appendix IV.

The second dataset, reported in section 5.2, consists of completed records of 460 actual test events, which are subjected to item response theory (IRT) analysis.

Subsequent sections outline how the methodology developed over the course of the project (5.3) and describe some possible alternative research methods (5.4).

5.1 Data set 1: Expert panel

The first data set was collected as part of a critical review of the Five Star test carried out in 1996 by Sheffield Hallam University TESOL Centre. The overall aim of this consultancy was to carry out a critical review of the pilot version of the test and to make recommendations for improving it.

The specific objectives that the critical review was to address were

1. Construct validation by skills analysis and difficulty rating
2. Calibration and reliability studies
3. Recommendations for test development
4. Content validation against external criteria tests
5. Development of tester orientation package

The source of data for objectives 1 - 3 was an expert panel based on the Delphi procedure, and this is discussed in detail in the following section. Objectives 4 and 5 were carried out by individual members of the panel after completion of the panel exercise, and these are discussed further in section 5.1.3 below.

The evidence collected by these activities was used as the basis of a report (Underhill, 1997) for the test developers, Saudi Development and Training (SDT), in January 1997 as the outcome of a consultancy project focusing on a critical review of the pilot version of the test. Considerable assistance was provided by SDT to facilitate this. The consultancy project was the first opportunity to submit the test to external scrutiny, and has been used among other sources to provide data for the redesigning and upgrading of the test to produce a commercially-available version. The focus of the report was therefore as much on proposals for test development as it was on test validation.

5.1.1 Expert panel: description of methodology

The Delphi method is a technique used for research and forecasting in social science, health and business to obtain the anonymous consensus of a group of experts on a complex question. It allows each individual to contribute on an equal basis and to draw on the evolving group consensus while precluding the bandwagon effect of dominant influence by one or two participants in a face-to-face discussion.

The technique is described in detail in Linstone and Turoff (1975) and more concisely in Delbecq et al. (1975) and Quade (1977). The introduction to Linstone and Turoff describes the origins of the Delphi technique in defence research, to

obtain the most reliable consensus of opinion of a group of experts ... by a series of intensive questionnaires interspersed with controlled opinion feedback... The original justifications for this first Delphi study are still valid today, when accurate information is unavailable or expensive, or evaluation models require subjective inputs to the point where they become the dominating parameters. A good example of this is in the "health care" evaluation area (Linstone and Turoff, 1975: 10)

The data from a series of Delphi panel rounds is used to support claims for content and construct validation, and as was seen in chapter three, these are types of validation which do not enjoy any simple or clearly-defined procedures. While experience suggests that individual judgments about the sampling or representativeness of items in a language test might vary quite widely, it is hypothesised that a consensus of language teaching professionals would emerge, and this hypothesis was tested by the analysis of data from Delphi panel rounds.

Linstone and Turoff identify a number of features that might lead to the need for employing Delphi:

- a) the problem does not lend itself to precise analytical techniques but can benefit from subjective judgments on a collective basis
- b) the individuals needed to contribute to the examination of a broad or complex problem have no history of adequate communication and may represent diverse backgrounds with respect to experience or expertise
- c) more individuals are needed than can effectively interact in a face-to-face exchange
- d) time and cost make frequent group meetings infeasible
- e) the efficiency of face-to-face meetings can be increased by ...group communication
- f) disagreements among individuals are so severe or politically unpalatable that the communication process must be refereed and/or anonymity assured
- g) the heterogeneity of the participants must be preserved to assure validity of the results, i.e. avoidance of domination ...('bandwagon effect') (Linstone and Turoff, 1975: 4)

In the current research, a) and g) are certainly true, while c) and d) provide practical reasons for employing such a technique. A further practical reason is that the

questionnaire for rounds 1-3 was completed at the keyboard and round 4 while watching video tapes, and it would be impossible to try to negotiate a face-to-face consensus among a panel of 12 members while simultaneously allowing each member to scrutinise a computer-based task or video segment.

Application of expert panel to language testing

In language teaching programmes, progress and achievement tests are routine events that are course specific, being drawn up for the most part by classroom teachers to reflect directly the recent content of the teaching syllabus. Being transitory and course- or class-specific in this way, no wider validity is sought or claimed, and informal consultation with peers and colleagues is a common way of pre-testing items where the time and resources do not exist to carry out a proper pilot programme. Such informal consultation is widely recommended by language testing authorities, e.g. Alderson et al. (1995: 63). However, there are few references in the literature to the formal use of an expert panel to make systematic judgments. Two examples are reviewed here: Lumley's (1993) investigation of reading subskills and Wijgh's (1993) validation of a framework for oral interaction.

The investigation of reading subskills (Lumley, 1993)

The existence of reading subskills as anything more than a working construct for teaching and research purposes, and their number, identity and hierarchy is a controversial topic, with intuitive analysis substituting for empirical research (Skehan

1984). Lumley (1993) used a group of five 'expert raters' in a complex decision making process to investigate which subskills are tested by particular reading test items. Panelists listed reading comprehension subskills in order of perceived difficulty, rated selected test items on the same scale, and then allocated the single most important skill needed to answer each question. "Judgements were compared and justified to the group..." (Lumley, 1993: 221) and a consensus was obtained in this way. This procedure differs crucially from the Delphi method used in the present research in that the negotiation after individual ratings had been made was conducted face-to-face, rather than anonymously with all the possible disadvantages listed above; in this context, the bandwagon effect being the most likely.

Two points that emerged from Lumley's study are reflected in the Five Star research. Firstly, panelists had considerable difficulty agreeing on the interpretation of the definitions of some of the subskills, and had to alter the wording of some; even then, there was still some potential confusion over hierarchy (whether one skill should be seen as included in another, or overlapping with it). The familiarisation round of the Five Star panel exercise described in the following section also allowed panelists to revise the skill definitions used.

Secondly, IRT was also used in the Lumley study to calculate item difficulty, which was then correlated against the subskill difficulty ratings given by the panelists. This analysis used QUEST, the same programme used for the current research (Adams and Khoo, 1996). The resulting significant correlation of .72 gives some empirical support to the validity of the panelists' collective judgements. Section 6.4.1 below reports on the use of correlation to explore the relationship between panel and IRT scores in the Five

Star test results, but in this case, the correlation is between two estimates of task difficulty.

The validation of a framework for oral interaction (Wijgh, 1993)

Wijgh (1993) used a panel of 12 members to explore the construct of oral interaction and to validate the framework for a test of oral interaction. She specifically invokes the Delphi procedure, and based on a sequence of steps in Harrell (1978), developed a two-stage procedure. Her first round invited panelists to comment on an initial working paper on the meaning and content of oral interaction ability, with a questionnaire asking them to agree or disagree with a series of statements based on the working paper. Open-ended comments were encouraged. These covered topics such as terminology and definitions, characteristics of language use, roles of the learner and types of interaction. A second questionnaire was developed based on the results of the first. Some original statements were left unchanged; some were reworded, entirely or slightly; some were omitted, but a question inserted seeking agreement for the omission. Some new statements were added, for example, about the use of language as a *lingua franca*. Her panel was deliberately chosen to represent a mixture of backgrounds:

As the framework was based on a synthesis of a literature study, scientific and theoretical knowledge had to be represented ... As the framework was developed for test construction, curriculum planners and test constructors had to be represented too. As the oral interaction ability is not a pure scholastic skill, but has to be used in real life,... some members of the panel had to represent the outer world, the commercial sector... (Wijgh, 1993: 5-6).

The resulting panel consisted of a teacher, two teacher trainers, a researcher in applied linguistics, a professor in linguistics, a professor in business communication, two curriculum planners, two test constructors and two managers. A high degree of consensus was reported amongst the panelists, and overall, the exercise was considered successful in validating the proposed framework for the oral test. The managers agreed

less with the group consensus than the others, possibly because they were the only members of the panel with no academic links. Panelists agreed highly on topics about language use and the role of the learner, but less highly on the topic of values in education; this topic seems to have been rather vague.

Wijgh reports a beneficial side-effect of the Delphi procedure was positive feedback from the experts on their participation, saying they had enjoyed and learnt from the experience, and an expression of willingness to continue their cooperation in the project in the future.

Other reference to expert panel procedures

McNamara also recommends the use of experts drawn from the target professional or vocational areas rather than restricted to English language teachers and testers: "People responsible for professional education and training for the workplace are likely to have a more explicitly articulated view of the nature of the workplace and its demands ... others such as work supervisors may also serve as informants." (McNamara, 1996: 94)

Alderson et al. recommend an expert panel for content validation in particular:

Typically, content validation involves 'experts' making judgements in some systematic way... An editing or moderation committee meeting ... may meet the requirements for content validation but only if the committee members can be considered to be experts, and if they are required to compare the draft test with its specifications or some other statement of content in a systematic way. In our experience this rarely happens. Instead, committee members opine on the content of items without much preparation, with no independent systematic attempt to gather their opinions, which means that the group dynamics of the committee meeting are likely to have a considerable influence on the outcome. (Alderson et al., 1995: 174).

It was to preclude this source of bias that the data collection procedure followed a series of rounds, with each panel member's views contributing equally and anonymously to the group consensus.

Alderson et al. (1995: 175) also advise caution in the expectation that experts will necessarily come up with the right answers, or even any consensus at all, creating a dilemma for the test developer who needs validity evidence as quickly as possible. The answer, they suggest (page 175), is not to abandon the expert opinion but to accept the complexity of language testing issues and to continue to gather evidence from diverse sources.

The American Psychological Association Standards also recommend the use of relevant experts external to the program for the test review process, and require that, however they are selected and used, "the qualifications, relevant experiences, and demographic characteristics of expert judges should also be documented". (APA, 1999: 44)

5.1.2 Expert panel: data collection

A panel of 12 experienced English language teachers and testers reviewed the Five Star test in a sequence of steps or 'rounds', working as an 'expert panel'. Panel members had from 5 to 25 years' teaching experience and practical experience each of at least one major ELT examination. The panel were divided at random into two groups, who were allocated alternate routes through the test in the familiarisation round and in all subsequent rounds.

The 73 working tasks in the test itself were divided among 4 routes a - d consisting of 3 routes of 18 working tasks each and one route of 19 tasks. Duplicate tasks inserted for

purposes of making the algorithm work were excluded, as were tasks originally identified as tests of writing. Allocation to routes was on number order basis, the first 18 working tasks being assigned to route a. and so on. The size of the routes was based on an estimate of 40-60 minutes as a comfortable concentration period allowing 2-3 minutes per task. Allocation to natural or authentic routes as experienced by test candidates following the test algorithm would have been ideal, but this was not practical for this exercise, as a very large number of natural routes would be necessary in order to cover all the cards.

In practice, the expert panel activities fell into three stages each with a distinct panel exercise.

Stage 1 consisted of preliminary orientation activities followed by a 'hands-on' panel exercise identifying which skills each of the 73 working tasks tested.

Stage 2 involved a similar 'hands-on' exercise asking panelists for a percentage allocation of each task to each skill and a determination of the minimum proficiency level necessary to complete each task successfully.

Stage 3 required panel members to view videos of completed tests in order to assign proficiency levels and to identify the verbal interaction strategies used by the candidates. At stages 2 and 3, panel members were also invited to make suggestions for improvements to individual tasks and to the test as a whole.

Each stage 1 -3 of the panel exercise is reported on separately below in text form, with *pro forma* copies of the data collection instruments used and the full statistical results given in the tables.

Expert panel stage 1: preliminary orientation activities and skills identification

The specific objectives at the first stage were:

- 1) to agree working definitions of language skills to be assigned
- 2) to familiarise panel members with the test and its individual tasks
- 3) to identify the major language skills tested by each task

Panel members carried out activity 1 as a paper exercise, and activities 2 and 3 individually at the keyboard.

Stage 1 description of procedure

The familiarisation round consisted of members of each sub-group working through the tasks in a different route, exploring all the various tasks, instructions and icons. This was felt to be essential before they were asked to begin making judgments, in the light of the novelty of computer-based adaptive tests in general and in particular Five Star's unique combination of adaptivity, live interaction and the computer platform. Panelists were subsequently presented with the working definitions of four skills, Listening, Speaking, Reading and Study skills shown in Table 7 and invited to comment on them. Their anonymous comments on these definitions were collated to produce revised definitions with between two and six alternative definitions for each skill, which were circulated again to panelists for them to select from. The aim here was to yield skill definitions which panel members had contributed to and were fully comfortable with. This consensus produced the final skill definitions used in subsequent rounds.

Five Star test validation : Language skills

We want you to assess cards on which skills they are testing, but first here are working definitions of four skills. Consider these definitions and say whether find them acceptable or unacceptable, and if you find them unacceptable, please suggest better definitions.

Listening

This skill requires the decoding and comprehension of samples of spoken English from the computer and/or the interlocutor, and placing them in context to complete a task that may require other skills also

acceptable / unacceptable If “unacceptable”, please give your definition:

Speaking

This skill requires the production of spoken English to complete a task that may require other skills also, in a way that the interlocutor can comprehend and evaluate against the criteria given

acceptable / unacceptable If “unacceptable”, please give your definition:

Reading

This skill requires the transliteration and comprehension of words, phrases, sentences and/or short texts on the screen and interpreting them in context to complete a task that may require other skills also

acceptable / unacceptable If “unacceptable”, please give your definition:

Study skills

These skills include numeracy (the ability to understand and manipulate numbers); transcoding (the ability to select, interpret and manipulate information from charts, diagrams and graphs); and instructions (the ability to understand and carry out task instructions given in English)

acceptable / unacceptable If “unacceptable”, please give your definition:

The original Five Star test algorithm allocates a participant's score across six skill factors, with Writing and Interaction in addition to the four listed above. The test clearly does not test Writing directly, and although theoretical arguments can be made for indirect testing, in practice the development of a direct writing component to add on to the existing test makes any review of this skill premature at this stage.

The reasons for omitting the Interaction skill from stages 1 and 2 are more complex. Interaction was considered:

- a) to be much harder to define than the other core skills, as discussed in chapter three, reducing the likelihood of achieving a workable consensus;
- b) necessarily to overlap substantially with listening and speaking, where the methodology theoretically demands independence between skills;
- c) to attach primarily to the event (the participants and the context of a particular test administration) rather than to the test item alone;
- d) to be premature at these stages. As the methodology adopted means that panelists were looking at the test tasks only in stages 1 and 2, and not observing video recordings of test administrations until stage 3, any attempt at assessment of interaction at stages 1 and 2 would be premature.

The first round of stage 1 allocated two of the four routes to each sub-group of six panelists. Each panelist was asked to complete the *pro forma* shown in Table 8, indicating whether they thought each task on each route tested each skill, didn't test it, or were unsure. 'No answer' was also scored as a fourth category of response. Additional comments were invited to qualify their responses.

Five Star test validation : skill allocation instructions

Task: show whether you agree with, disagree with or are unsure about the statement that these cards test each of these skills. Please make a note under "other comments" if you feel your response should be qualified or the card assesses any other skill. If necessary, continue over the page.

- Listening This skill requires the decoding and comprehension of samples of spoken English from the computer and/or the interlocutor, and the placing of them in context to complete a task that may require other skills also.
- Speaking This skill requires the production of spoken English to complete a task that may require other skills also, in a way that the interlocutor can comprehend and evaluate against the criteria given.
- Reading This skill requires the transliteration and comprehension of words, phrases, sentences and/or short texts on the screen and their interpretation in context to complete a task that may require other skills also. As this test is designed for native speakers of Arabic, 'transliteration' here refers to the process of decoding the characters of an unfamiliar alphabet.
- Study skills These skills include numeracy (the ability to understand and manipulate numbers); transcoding (the ability to select, interpret and manipulate information from charts, diagrams and graphs); and following instructions (the ability to understand and carry out task instructions given in English)

	listening	speaking	reading	study skills	other comment
Card 4	yes/no/unsure	yes/no/unsure	yes/no/unsure	yes/no/unsure	
Card 5	yes/no/unsure	yes/no/unsure	yes/no/unsure	yes/no/unsure	
Card 6	yes/no/unsure	yes/no/unsure	yes/no/unsure	yes/no/unsure	
Card 7	yes/no/unsure	yes/no/unsure	yes/no/unsure	yes/no/unsure	
Card 8	yes/no/unsure	yes/no/unsure	yes/no/unsure	yes/no/unsure	
Card 10	yes/no/unsure	yes/no/unsure	yes/no/unsure	yes/no/unsure	
Card 11	yes/no/unsure	yes/no/unsure	yes/no/unsure	yes/no/unsure	
Card 12	yes/no/unsure	yes/no/unsure	yes/no/unsure	yes/no/unsure	
Card 13	yes/no/unsure	yes/no/unsure	yes/no/unsure	yes/no/unsure	

The second round of stage 1 repeated the process, with each group of panelists seeing the other two rounds, in the pattern shown in Table 9. The purpose of this design was to counter any consistent bias caused by the order in which panelists were exposed to task.

Table 9 **Routes through Five Star tasks for panel stage 1**

	<i>group 1</i>	<i>group 2</i>
<i>round 1</i>	routes a & b	routes c & d
<i>round 2</i>	routes c & d	routes a & b

Thus each panelist has expressed in stage 1 a 'yes/no' judgement on whether or not each skill is tested by each task in the test. The resulting total of some 3500 judgements have been collated to produce the stage 1 results. Where a consensus was achieved, this provides some evidence of the construct validity of each task.

Expert panel stage 2: percentage skills allocation and determination of proficiency level

The specific objectives at the second stage were:

- 1) to allocate percentages of language skills to each task
- 2) to identify a minimum level of proficiency needed to perform each task
- 3) to make suggestions for improving each test task

Panel members carried out these activities individually at the keyboard. This stage built on the gradual exposure of panel members to the test tasks in stage 1 and combined the functions of validation activity and critical review for test development purposes. The consensus on skills allocation and proficiency levels has been used in the preparation of the next version of the test to improve the algorithm that underlies the computer routing of each test administration through the tasks.

Stage 2 description of procedure

In the first round of stage 2, the same panel members in the same two sub-groups worked through the same routes at the keyboard to allocate a percentage range of each task to each relevant skill and the minimum level on an external scale needed to complete each task successfully. For each task, the pro forma shown in Table 10 asked the panelists for suggestions on improving the content and the presentation of the task separately. The second round of stage 2 repeated the process, with each group of panelists seeing the other two rounds, in the same pattern as in stage 1, shown in Table 9. Thus each panelist expressed in stage 2 a numerical judgement on the extent to which each task tested each skill, and on the proficiency level needed to complete it successfully.

The use of a percentage range for skill allocation, going up in steps of 5% from 0-4%, 5-9%, etc, allowed panelists to make a sufficiently fine allocation for each skill for each task without asking for the unrealistic over-precision of a discrete percentage score. At the same time, it allowed greater freedom for the panelist to express percentage ranges totalling less than 100%, in other words, to imply that there were other significant factors contributing to success with a particular task than the four skills being allocated.

Card	listening % code A-V	speaking % code A-V	reading % code A-V	study skills % code A-V	Min. level 1 - 9

How can the **content** of this card be improved?

How can the **presentation** of this card be improved?

Card	listening % code A-V	speaking % code A-V	reading % code A-V	study skills % code A-V	Min. level 1 - 9

How can the **content** of this card be improved?

How can the **presentation** of this card be improved?

Table 10 (continued) Panel stage 2 *pro formas*: language skill and level allocations

Language proficiency levels

- Level 9 Has a full command of the language, tackling the most difficult tasks with consistent accuracy, fluency, appropriate usage, organisation and comprehension. An exceptional level of mastery, not always reached by native speakers, even quite educated ones.
- Level 8 Uses a full range of language with proficiency approaching that in the learner’s own mother tongue. Copes well even with demanding and complex language situations. Makes occasional minor lapses in accuracy, fluency, appropriacy and organisation which do not affect communication. Only rare uncertainties in conveying or comprehending the content of the message.
- Level 7 Uses language fully, effectively and confidently in most situations, except the very complex and difficult. A few lapses in accuracy, fluency, appropriacy and organisation, but communication is effective and consistent, with only a few uncertainties in conveying or comprehending the content of the message.
- Level 6 Uses the language with confidence in moderately difficult situations. Noticeable lapses in accuracy, fluency, appropriacy and organisation in complex situations, but communication and comprehension are effective on most occasions, and are easily resorted to when difficulties arise.
- Level 5 Uses the language independently and effectively in all familiar and moderately difficult situations. Rather frequent lapses in accuracy, fluency, appropriacy and organisation, but usually succeeds in communicating and comprehending the general message.
- Level 4 Uses a basic range of language, sufficient for familiar and non-pressuring situations. Many lapses in accuracy, fluency, appropriacy and organisation, restricting continual communication and comprehension, so frequent efforts are needed to ensure communicative intention is achieved.
- Level 3 Uses a limited range of language, sufficient for simple practical needs. In more exacting situations, there are frequent problems in accuracy, fluency, appropriacy and organisation, so that normal communication and comprehension frequently break down or are difficult to keep going.
- Level 2 Uses a very narrow range of language, adequate for basic needs and simple situations. Does not really have sufficient language to cope with normal day-to-day, real-life communication, but basic communication is possible with adequate opportunities for assistance. Uses short, often inaccurately and inappropriately worded messages, with constant lapses in fluency.
- Level 1 Uses a few words or phrases such as common greetings, and recognises some public notices or signs. At the lowest level, recognises which language is being used.

Table 10 (continued) Panel stage 2 *pro formas*: language skill and level allocations

Five Star test validation : skill and level allocation instructions

Task

- using the code letters A - V below, indicate the percentage range to which you think each of the cards tests each of the skills
- indicate the minimum language proficiency level, from 1 to 9 on the scale overleaf, needed to complete the task successfully
- suggest how the card could be improved, either for content or for presentation. If necessary, continue over the page.

Percentage ranges:

0 - 4%	5 - 9%	10- 14%	15- 19%	20- 24%	25- 29%	30- 34%	35- 39%	40- 44%	45- 49%	50- 54%	55- 59%	60- 64%	65- 69%	70- 74%	75- 79%	80- 84%	85- 89%	90- 95%	96- 100%
A	B	C	D	E	F	G	H	J	K	L	M	N	P	Q	R	S	T	U	V

Skill definitions

Listening

This skill requires the decoding and comprehension of samples of spoken English from the computer and/or the interlocutor, and the placing of them in context to complete a task that may require other skills also.

Speaking

This skill requires the production of spoken English to complete a task that may require other skills also, in a way that the interlocutor can comprehend and evaluate against the criteria given.

Reading

This skill requires the transliteration and comprehension of words, phrases, sentences and/or short texts on the screen and their interpretation in context to complete a task that may require other skills also. As this test is designed for native speakers of Arabic, 'transliteration' here refers to the process of decoding the characters of an unfamiliar alphabet.

Study skills

These skills include numeracy (the ability to understand and manipulate numbers); transcoding (the ability to select, interpret and manipulate information from charts, diagrams and graphs); and following instructions (the ability to understand and carry out task instructions given in English)

Table 10 Panel stage 2 *pro formas*: language skill and level allocations

The external scale

The external rating scale used was the nine-level Overall Language Proficiency scale, and this is given in full in Table 11. This was a product of the English Speaking Union's 'Yardstick' project culminating in the publication of the ESU Framework document (Carroll and West, 1989).

Table 11 External rating scale used for panel judgements of proficiency

Level 9	Has a full command of the language, tackling the most difficult tasks with consistent accuracy, fluency, appropriate usage, organisation and comprehension. An exceptional level of mastery, not always reached by native speakers, even quite educated ones.
Level 8	Uses a full range of language with proficiency approaching that in the learner's own mother tongue. Copes well even with demanding and complex language situations. Makes occasional minor lapses in accuracy, fluency, appropriacy and organisation which do not affect communication. Only rare uncertainties in conveying or comprehending the content of the message.
Level 7	Uses language fully, effectively and confidently in most situations, except the very complex and difficult. A few lapses in accuracy, fluency, appropriacy and organisation, but communication is effective and consistent, with only a few uncertainties in conveying or comprehending the content of the message.
Level 6	Uses the language with confidence in moderately difficult situations. Noticeable lapses in accuracy, fluency, appropriacy and organisation in complex situations, but communication and comprehension are effective on most occasions, and are easily resorted to when difficulties arise.
Level 5	Uses the language independently and effectively in all familiar and moderately difficult situations. Rather frequent lapses in accuracy, fluency, appropriacy and organisation, but usually succeeds in communicating and comprehending the general message.
Level 4	Uses a basic range of language, sufficient for familiar and non-pressuring situations. Many lapses in accuracy, fluency, appropriacy and organisation, restricting continual communication and comprehension, so frequent efforts are needed to ensure communicative intention is achieved.
Level 3	Uses a limited range of language, sufficient for simple practical needs. In more exacting situations, there are frequent problems in accuracy, fluency, appropriacy and organisation, so that normal communication and comprehension frequently break down or are difficult to keep going.
Level 2	Uses a very narrow range of language, adequate for basic needs and simple situations. Does not really have sufficient language to cope with normal day-to-day, real-life communication, but basic communication is possible with adequate opportunities for assistance. Uses short, often inaccurately and inappropriately worded messages, with constant lapses in fluency.
Level 1	Uses a few words or phrases such as common greetings, and recognises some public notices or signs. At the lowest level, recognises which language is being used.

'Yardstick 1 Stage I : Overall language proficiency' (Carroll & West, 1989:21)

The major advantage of this scale is that it was specifically designed to enable comparisons between different English language tests and examinations, and was not tied to one particular test or examining board. The disadvantage is that its validation rests largely upon the *a priori* research that went into its development, and there has been no systematic attempt to validate it independently. However, it is widely used by examining boards and other test developers, and it is its uniquely non-partisan orientation that makes it an appropriate choice for the validation and critical review of a new test.

The Overall Language Proficiency scale is the most general of a total of 22 scales in the ESU Framework, the others being related to specific skills and specific purposes, such as 'speaking for business purposes' or 'writing for study/training purposes'. Where one purpose of the current exercise is precisely to determine which skills are being tested, it would be anticipating the results to use skill-specific scales for the purpose. Even after the preliminary skills identification, such scales would still be inappropriate where the majority of tasks were found to be testing at least two of the major skills.

Expert panel stage 3: scoring video tests and identifying verbal interaction strategies

The specific objectives at the third stage were:

- 1) to assign an overall level of proficiency to each of eleven candidates on video tape.
- 2) to identify verbal interaction strategies other than the core language skills which affected the overall performance
- 3) to elicit further suggestions for improving the test

Panel members carried out these activities individually viewing video tapes of authentic test events. This yielded reliability data comparing the judgements of panel members against each other and against the existing test procedure, and started to calibrate the test against an external scale.

Stage 3 description of procedure

The same panel members allocated to the same two groups viewed two videotapes containing a total of 11 recordings of authentic administrations of the Five Star test, lasting between 10 and 40 minutes. Group one watched tape A then tape B, group two the other way round. Panelists worked through a rubric which listed the tasks occurring in each test event, and for each task in each test they indicated which verbal interaction strategies they noticed being used, from a pre-set list of six with an additional column for 'other' strategies. They were given the examples of each interaction strategy shown in Table 12.

Table 12 Panel rubric for interaction strategy exercise

Column number	Interaction strategy	Example exponents
1	confirms understanding	'I understand', 'yes/yeah', agrees, disagrees, laughs, ...
2	seeks confirmation	'Do you mean...?', repeats with question intonation,...
3	seeks clarification	'I don't understand', 'I'm sorry', 'please repeat',...
4	indicates need for clarification	Fails to respond, extended silence, 'errr...', 'ummm...', ...
5	confirms own previous turn	'yes', 'that's right', or equivalent
6	re-forms own previous turn	'no, I meant ...', rephrases previous statement, ...
7	Other:	Any other verbal interaction strategy

For each video test event as a whole, they then indicated whether they felt the interaction strategies contributed to or detracted from the candidate's performance; identified any other features or skills other than the interactional strategies and core

skills, which influenced the outcome of the test event; and gave an overall proficiency rating for that candidate's performance, using the same ESU scale as at stage 2. Finally, panelists had a final opportunity to make suggestions for improving the existing tasks or topics; adding new tasks or topics; and any other general suggestions or recommendations.

5.1.3 Content validation against external criteria tests and development of tester orientation package

In addition to the expert panel exercise described above, the specification (section 5.1) for the critical review (Underhill, 1997) included two other activities: the content validation of Five Star against external criteria tests and a framework for the development of an assessor orientation package. These activities were carried out by individual members of the panel after completion of the panel exercise, for two reasons. Firstly, the authors had spent a considerable amount of time contributing to the panel exercise, including scrutiny of every task and viewing of videos of test events, so that they were fully familiar with the test before carrying out the comparisons or writing the assessor training framework; and secondly to prevent these activities creating a differential source of bias that might give panelists different perspectives on the test during the panel exercise.

Content validation against external criteria tests

Examiners drawn from the expert panel made content comparisons against three established examinations with a direct oral component: the UCLES 'main suite' EFL

exams, the Trinity College London Spoken English Grade Exams and the IELTS test (these three were included in the summary comparison against other oral test formats in Table 6 in chapter four). The authors were current or former accredited examiners of the criterion test they were comparing. These comparisons sought to provide evidence for content validity and construct and, in the absence of empirical data allowing concurrent scores to be compared, some evidence for criterion validity. These examinations were chosen for comparison because

- a) they all contain a component of direct oral testing
- b) they all contain at least some features of the communicative approach to proficiency testing and between them represent a good range of current oral testing practices
- c) at the same time, they are all undertaken as commercially-viable activities, and must therefore have addressed some of the issues faced by the Five Star test in the operationalisation of its approach
- d) they are all available in Saudi Arabia.

These content comparisons against external tests looked at the different language skills tested; the specific techniques used for elicitation of the spoken language sample; the topics and patterns of interaction employed; and the scoring or measurement systems used. The authors were given these as possible areas for discussion but were not required to follow a specific reporting format; they were rather encouraged to bring out what they felt emerged as the most interesting features of the comparison. In each case, there are similarities and differences; these are discussed below in section 6.2 and are given in full in Appendix IV (McCarthy, 1997; Graham, 1997; Kontoulis, 1997).

Development of tester orientation package

The critical review report also contained a preliminary framework for training new test administrators, in recognition of the fact that the administration of almost all the pilot version tests by a single administrator was both a source of consistency for the pilot version data and a potential source of bias when delivered by others. Wider commercial availability of the test will necessarily involve other test assessors, for whom a training programme was essential to ensure consistency of delivery of the test between and across assessors.

These proposals were prepared by a member of the expert panel with significant experience both in English language teacher education and the management and standardisation of tests of spoken language. It is quoted in full in Appendix V: 'A framework for training and licensing Five Star assessors' (Parker, 1997).

The proposals describe the outline of a process for licensing assessors, and fall into three stages: Pre-training, Training and Post-training. The Training component is further broken down into theoretical and procedural induction, practical induction and the training manual. No separate data are available to describe the pre- and post-training procedures implemented by the test publisher for new assessors, but an assessor's manual has been produced in the last year (SDT, no date). The manual draws on key features of the framework proposal given in Appendix V, and incorporates many of its recommendations such as a list of assessor selection criteria and the outline of an induction course as well as a background to the test, detailed instructions for administering it and discussion of aspects of test theory.

5.2 Data set 2: Item response theory

5.2.1 Item response theory: description of methodology

The background to IRT is described in chapter three above, where the evolution of a broader class of item response models was mentioned. One of these variations enables the analysis of polytomous variables, in other words, test items where there are more than two possible outcomes, right or wrong. The Five Star test has three exit buttons on the screen at the bottom of each task, and the interviewer ends a task by deciding which of the three labels (see column 4 'scoring criteria' in Table 4 above) most closely describes the candidate's performance and clicking on the appropriate button. Putting it crudely, a choice has to be made between low, medium and high performance on each task, and this is known as a 'partial credit' model.

The more sophisticated two- and three-parameter IRT models can only be used on dichotomous data, and are therefore unsuitable for Five Star data. In practice, the absence of the second and third parameters is not greatly significant; the second parameter, item discrimination, is reflected indirectly by the item fit statistics generated by the one-parameter model anyway, and because of the subjective evaluation of performance for most tasks there are few opportunities in the Five Star test where the third parameter, guessing, could operate significantly. The guessing parameter is particularly appropriate to true/false and multiple choice items (Baker, 1997: 58)

The QUEST software (Adams and Khoo, 1996) used for the analysis of this data is therefore a one-parameter model based on Masters (1982) and Wright and Masters (1982), where 'partial credit' is distinguished from three other types of ordered category data: repeated trials, counts and rating scales. A brief outline of these four types may be useful.

What these four research designs share in common is that they record an ordered level of response for each subject or candidate, rather than single 'right or wrong' outcomes. The common algebraic form gives the probability of a subject of ability β scoring x rather than $x-1$ on an item of difficulty δ . This is a special case of the basic Rasch model for dichotomous data which gives the probability of the subject of ability β getting an item of difficulty δ right rather than wrong (Wright and Masters, 1982: 55).

The first type, *repeated trials*, "result when respondents are given a fixed number of independent attempts at each item on a test. The observation x is the number of successes on the item ... the order in which success occur is considered irrelevant". The second type, *counts*, "results when there is no upper limit on the number of independent success (or failures) a person can make on an item'. The third type, *rating scales*, "comes from rating scales in which a fixed set of ordered response alternatives is used with every item". The fourth type, *partial credit*, "comes from an observation format which requires the prior identification of several ordered levels of performance on each item and thereby awards partial credit for partial successes on items" (Masters, 1982: 150)

The rating scale and partial credit models are quite similar. The crucial difference is that the rating scale model requires a single set of response categories for every item in the scale, whereas the partial credit model is "an extension of Andrich's Rating Scale model to situations in which response alternatives are free to vary in number and structure from item to item" (Masters, 1982: 150). In other words, the rating scale model requires that the steps between the response categories be uniform across all the items in the test; for example, that the difference that is being measured in a questionnaire between 'Agree' and 'Strongly agree' must be the same for all the items. "The relative difficulties of the steps in a rating scale item are usually intended to be governed by the fixed set of rating points accompanying the items. As the same set of rating points is used with every item, it is usually thought that the relative difficulties of the steps in each item should not vary from item to item." (Wright and Masters, 1982: 48)

By contrast, "The partial credit model imposes no particular expectations concerning the relative difficulties of the steps within any item. Some will be easy, others hard, depending upon the item subtasks, and quite regardless of the necessary order in which the steps must always be taken." (Masters, 1982: 162) Although in the case of some Five Star tasks, the scoring criteria are similar to a three-step rating scale, the test constructors made no claim for the similarity of step structure and difficulty across tasks, and given the diversity of criteria and score systems used for different tasks, it would not be possible to justify such an assumption. The partial credit rather than the rating scale model of IRT is therefore appropriate.

Two assumptions underlying the IRT model referred to in section 3.5, unidimensionality and local independence, should be considered in the context of the Five Star test.

Unidimensionality requires that the attribute underlying performance should be the same for all items in a test. There has been considerable discussion in the literature (e.g. Baker; 1997; McNamara, 1996; Crocker and Algina, 1986) as to how strictly this can be interpreted, and indeed whether it can even be measured. Unlike classical item analysis which typically also assumes unidimensionality, IRT generates an indicator of the extent of violation of this assumption, in the form of fit statistics for each item. In the case of the Five Star test, the developers' original assumption, reflected in both the algorithm and the scoring system, was that there were six different language-related skills involved, in different combinations and proportions. The skills allocation activities in the expert panel exercise described in 5.1.2 above carried over this assumption, but based on four rather than six skills, and succeeded in reaching a high degree of consensus about the relative skills allocation for most of the tasks.

It would therefore be possible in principle to hypothesize a 'separate skills' model for the Five Star test which could be tested out through a re-analysis of IRT data postulating these skills as a distinct test facet. However, the original construct was based very much on tasks requiring the display of integrated skills as a reflection of authenticity. In real life, we do not often use only speaking or listening, but typically both, and perhaps other skills at the same time; and evidence for this integrated skills model is provided by the panel consensus that most tasks required two or more of the skills allocated (section 6.1.1 below). On this basis, therefore, seeking evidence for separate language skills as

non-overlapping multiple dimensions would seem to be just as artificial an exercise as assuming a single unidimensional skill. An alternative argument for a unidimensional approach would simply be that there is an underlying language proficiency construct which is accessed through varying modes of communication, and additionally all the tasks in the test have in common an element of live interaction between candidate and interviewer.

An experiment to find out whether a language test battery with sub-tests of different skills violated the assumption of unidimensionality was carried out by Henning et al. Over 300 students took a placement examination consisting of 150 multiple choice items, 30 in each of five sub-tests designed to assess Listening Comprehension, Reading Comprehension, Grammar Accuracy, Vocabulary, and Writing Error Detection. The hypothesis that each sub-test was itself a separate content domain that would violate the assumption of unidimensionality of the test as a whole was tested by comparing the Rasch-generated difficulty scores for each sub-test against the test as a whole. Additionally, the number of misfitting items was compared for each sub-test against the analysis of those items in the test as a whole. They found no violation of unidimensionality and concluded that

item response theory in general, and Rasch Model in particular, are sufficiently robust with regard to the assumption of unidimensionality to permit applications to the development and analysis of language tests which may be comprised of item domains representing diverse subskills of language use... (Henning et al., 1985: 152)

Such an experiment was possible because all the items in their test battery fell neatly into discrete sub-tests. It would be harder to carry over this analysis to Five Star test results because the task design deliberately uses tasks that tap integrated skills. Given the similarity between the type of sub-skill involved it seems reasonable to believe that

their conclusion is valid for Five Star also, and that the Rasch model is robust with respect to unidimensionality.

Local independence requires that items are not interdependent, so that success or failure on one item is not directly linked to performance on another. Examples of language tests where this can be problematic are cloze tests where understanding of parts of a text may influence the ability to complete non-local gaps, i.e. to supply words that have been deleted earlier or later in the passage. Similarly, some reading or listening comprehension tasks have a series of questions that all relate to the same text; successful comprehension of the text may influence success or failure on the series of items, so that performance on each item in the series is not independent of the others. This is why the reading section of the new computer-based TOEFL test is not adaptive, see section 4.5 (iii) above.

In theory, the task-based approach adopted by the Five star test ensures local independence, with each task completed and scored before the next task is presented and attempted. However, there is a conflict here with the construct of interaction conceived as a strategic skill operating over a series of tasks (section 4.2). The developer sees the possibility of a topic being pursued or developed over a series of tasks as a positive feature (Pollard, 1998a) that authentically reflects real-life behaviour. Any significant evidence that the interaction between candidate and interviewer that is unique to each test event was carried over and developed from task to task might be taken to undermine the assumption of local independence. There is however no deliberate carry over of information from one task to the next, and the adaptive nature of

the test creates a very large number of different permutations and sequences of test tasks, making the possible effect difficult to research.

A second difficulty with the assumptions is precisely that they are ultimately unverifiable. Unidimensionality and local independence are in fact linked, to the extent that items calling on the same underlying trait will in fact have some interdependence, and one definition of a latent trait is that it accounts for the statistical interdependence between items. "Local independence and the number of latent traits are always a matter of assumption." (Crocker and Algina, 1986: 343)

5.2.2 Item response theory: data collection

After deletion of a small number of duplicates and defective records, there were 460 complete test records available for analysis. The Five Star test produces an optional printout which records each score on each of the tasks tackled by the candidate, together with the time taken, date of the test event and the scores on the form of the graphic Five Star profile as well as the scores calculated for the six sub-skills. These tests events all took place in Saudi Arabia between January 1994 and November 1997, and were all conducted by the same person.

The data from the printed score outputs were transferred to a single spreadsheet which shows all the scores on all the items taken by all the candidates. A simple count identifies how many candidates took each item, and what scores were awarded; how many of the candidates for each item were assessed at the low, medium and high

performance scores. These are effectively the item scores, and in a conventional test where all the candidates take all the items it would then be possible to calculate two of the most useful classical test statistics:

- the item facility, by seeing what proportion of the total sample was successful on the item
- the item discrimination, by comparing how many of the candidates who scored high overall performed on a given item compared to the candidates who scored low overall.

However, because of the adaptive nature of the test, it is impossible to make any direct comparison between these item scores or to produce conventional item facility or discrimination indices. The fact that, say, fifty out of a hundred candidates succeeded on one task compared to only thirty out of eighty on another cannot be interpreted to suggest that the first task is easier than the second, as the sample of candidates may have been significantly different. Indeed, they may have been taken to different tasks by the algorithm precisely on the basis of differential performance on earlier tasks.

All that can usefully be inferred from the raw table of scores is that certain items are very rarely used, for example, there are several that occur less than 10 times in 460 test events. In a revised version of the test either the algorithm would need to be redesigned to use them more frequently or their inclusion justified on the grounds of comprehensiveness of coverage of level and skill.

The spreadsheet of scores was converted to a text file and subjected to analysis using QUEST (Adams and Khoo, 1996), a computer program designed for IRT analysis of

partial credit data. A series of outputs from the QUEST analysis report task difficulty and candidate ability levels together with standard error measurements and goodness of fit statistics for each. These are described in more detail in section 6.3.

5.3 How has the methodology developed over the project?

The project started with a critical review of the Five Star pilot test, carried out for the test developers in 1996, and described in section 5.1 above. The specific requirement for recommendations for test development entailed substantial qualitative judgement in the review process. Some of the objectives such as calibration and reliability studies could have been met with a purely empirical analysis, had sufficient data been available, and construct validation could in theory have been approached through the kind of procedures described in section 3.2.3. Similarly, criterion validity could have been based in part on concurrent comparison of test results on Five Star and external tests.

At that time, when The Five Star pilot test had only been in operation for two years, only about 200 completed test records were available. This was not enough for IRT analysis, given that each candidate typically only attempts 10-15 of the test tasks, nor was the incomplete data set generated by the adaptive nature of the test susceptible to other psychometric approaches (see section 5.4 below). No scores had been obtained on external criterion tests.

An expert panel approach therefore seemed the most appropriate method of analysis that allowed both decisions based on empirical data such as difficulty level and skills

breakdown and at the same time generated a rich source of open-ended suggestions and recommendations for improving the test and tasks. Given the complexity of some of the decisions, it was agreed that a Delphi procedure would be the best way to minimise the dangers of peer influence and the bandwagon effect.

At the same time, the familiarity gained by individual experts through the panel exercise was utilised by inviting them to draw up comparisons of the Five Star test against external exams with which they had been professionally involved, thus addressing the objective of content comparison in the absence of criterion test scores.

Among issues raised by the critical review was the difficulty of carrying out a validation exercise on a test at an early stage of development when little empirical data has been collected, and the paucity of explicit models in the literature for such validation. However, it is precisely at this stage that solid evidence is needed to make crucial decisions about the future development of the test. In the case of Five Star, the combination of a number of attributes, such as adaptivity, the computer platform, the implications of live oral interaction and other features of the communicative methodology, emphasized the unique and ephemeral nature of each test event. Taken with recent changes in approaches to validity, these test features exacerbated the difficulty of establishing validity for such a test as a once-and-for-all status by traditional means.

After the submission of the critical review, the scope of the project was subsequently extended to the current research and the question of how validation for such a class of tests might be approached became the central focus. Further empirical data was clearly

needed, and IRT is widely recognised as the most suitable candidate for the analysis of adaptive test data (section 3.5 above). By the end of 1997, a total of 460 completed test records were available, which was sufficient for a preliminary IRT analysis. As well as generating a second large dataset this also made possible the comparison of results with the expert panel analysis of the test (section 6.4 below).

During the process of developing a systematic validation model, the constructs underlying the test needed clarification in the absence of a very explicit statement in the Five Star test literature. Scrutiny of the test suggested that it fits well with the current communicative approach to language testing, but goes further in applying many of the features of communicative testing to a computer-based delivery system; this is apparently unique. Investigation of these features and how they are evidenced in the test has contributed not only to an overall judgement on the Five Star test validity, but in a significant way to the necessity of constructing the validation model as a continuing process. It is in the essence of communicative attributes such as authenticity and topicality that they are heavily context-dependent, in time, place and social context; and it is the status and inter-relationship of features such as interaction, authenticity, individualisation and directness that has emerged as an enduring problem for the construction of the validation model.

5.4 Alternative research methods

Ethnomethodological approaches

Ethnomethodological approaches relying on feedback from test subjects could provide first-hand evidence of how they interpret each task and the thought processes they bring to it. A variety of techniques have been well documented for eliciting and analysing introspective first-person accounts (Cohen and Manion, 1994: 204) This would be very useful for supporting evidence of which skills different tasks were testing, and so indirectly for construct validity, but suffers from the usual disadvantages of face validity, that is may be based on superficial judgements only for which no further supporting evidence can be found. In an adaptive test in particular, it also means that each candidate is only able to comment on a small proportion of the total number of items. While it might provide illuminating information on individual tasks, it would undermine the test developers' aspiration to create and validate a more conversational interaction that carries over a series of tasks. The same criticism is however true for any research method, such as the IRT employed in this study, that is based on the individual task as the unit of analysis.

To prevent language proficiency being a major constraint on the quantity and quality of data generated by such approaches, the elicitation procedures would have to be carried out in the candidates' mother tongue. This would be much easier in the case of a language test such as Five Star designed for use in a monolingual environment than for a multinational context, but there would still be considerable practical problems surrounding the collection of data.

Discourse-analytic approaches

Research methods based on an analysis of the actual language used by subjects is particularly relevant to language teaching and testing as it is the focus of the whole enterprise as well as the medium for eliciting research information. Originally applied to written language, discourse analysis now subsumes interaction analysis and conversation analysis as sub-types.

Such analyses of the language used in spoken tests could generate valuable information at two levels. Specifically, they could throw light on the discourse components variously labeled as pragmatic, textual, rhetorical or rules of discourse in the componential models described in section 2.9.2 which would help determine whether these were observable traits that could usefully be included in a rich model of oral testing and whether they contributed to a measure labeled 'interaction'.

More generally, they could provide detailed evidence to compare the actual linguistic and para-linguistic behaviour in the test event with the target language use, and so provide specific evidence for or against content and construct validity of the test in general. To make this possible, parallel discourse-analytic studies would need to be made of the situations of target language use. Alternatively, they might indicate that the language of tests was more like the language of the classroom (van Lier, 1996) than real life conversation.

Some frameworks for analysing oral interaction, and conversation analysis in particular were considered in section 3.3.2. Pollard (1999) has started to analyse excerpts from

Five Star test recordings, comparing at one level assessor to candidate eye contact as well as measures of length and number of spoken turns (1998a) and at another level a very detailed transcription, including length of pauses, amplitude of voice and some non-verbal gestures.

The major barrier to widespread use of discourse-analytic approaches is that they are immensely time-consuming, requiring many hours of painstaking analysis and transcription for a few minutes of a single test recording. This immediately raises questions about the representativeness of the candidates and the excerpts chosen for analysis which may make it difficult to generalise any conclusions drawn to the validity of the test in general.

Psychometric approaches

Factor analysis (Kline, 1993: 93) attempts to account for correlations between scores by identifying patterns of variation which can be associated with different factors, which are hypothetical constructs estimated from the data. There are potentially an infinite number of factors, and many different types of factor analysis, but the ultimate aim is to explain the variance that is observed with reference to variables which might be hypothesized in advance (confirmatory factor analysis) or which might emerge from the data (exploratory factor analysis). However, like analysis of variance, factor analysis cannot be performed on incomplete datasets of kind generated by adaptive tests, and therefore could not be used to analyse test data in the current research.

Concurrent validity

Concurrent validity requires the comparison of test results against scores on a criterion measure (Bachman, 1990: 248). As the Five Star is an entirely new test, no criterion measures have yet been obtained. When the new version of the test is fully operational, comparisons against other English language proficiency tests such as TOEFL, IELTS or Cambridge examinations can be obtained, but correlations between them will have to allow for the influence of features such as the local specificity and computer-based delivery. Exceptionally high correlation would be suspect as it would imply that there was no effective difference in what the tests were measuring; it would also suggest that the new test was redundant if it correlated so highly with an existing one.

Although empirical concurrent correlation has not been possible, the content comparisons (see sections 5.1.3.1, 6.2 and Appendix IV) in the critical review against three other major established tests provide some evidence of concurrent validity.

Predictive validity

The Five star test benefits from having clearly defined contexts for target language use, and it would in theory be feasible to collect appropriate data. Predictive validity criteria (Bachman, 1990: 250) might be individual performance on future job-related tasks or achievement in English-medium training, or surveys of managers, tutors or co-workers to establish the individual's success in communicating through English. Because of the likely time lapse between test result and performance on criterion behaviour, interpretation of correlations for predictive validity must take account of a range of intervening factors which cannot be controlled, such as varying exposure to English language. Such a project would be prohibitively expensive on anything other than a very small scale and might very well be inconclusive.

5.5 Summary

This chapter has outline the methodology and data collection for the two datasets used in this study, and has illustrated how they were linked over the period of the project and what research methods were considered.

From two entirely different sources, the two datasets can be seen as complementary. The expert panel consists of a very large number of individual subjective judgements, but brought together in a carefully designed research plan to optimise the opportunity for each panelist to contribute equally while minimising the possibility of a bandwagon effect. It can take place during the planning or pilot stage of test development, before sufficient test events have been completed for a detailed post hoc analysis of test.

Precautions taken against bias in the panel activities included

- the choice of the Delphi procedure, with panelists taking part anonymously, even those who were in regular daily contact
- the random allocation of panelists to two groups who took different routes through the test
- the familiarisation round to give all panelists an equal orientation to the Macintosh computer and test software
- the opportunity for panelists to comment on skill definitions which were subsequently reformulated and circulated again for further comments before use

The IRT analysis was a completely objective treatment of actual test event data, but for that reason can only be carried out when sufficient test have been completed for which records are available. Issues raised by the assumptions of unidimensionality and the number of latent traits remain unresolved but the literature suggests that the Rasch model is sufficiently robust to be used in such a situation.

- 6.0 Introduction
- 6.1 Data set 1 : expert panel
 - 6.1.1 Expert panel stage 1 results and analysis
 - 6.1.2 Expert panel stage 2 results and analysis
 - 6.1.3 Expert panel stage 3 results and analysis
- 6.2 Data set 1 : content comparisons
- 6.3 Data set 2: IRT results and analysis
 - 6.3.1 IRT outputs 1: summary statistics for items and cases
 - 6.3.2 IRT outputs 2: individual statistics for item estimates
 - 6.3.3 IRT outputs 3: individual statistics for case estimates
 - 6.3.4 IRT outputs 4: distribution of both items and cases
 - 6.3.5 IRT outputs 5: individual case (candidate) maps
- 6.4 Comparison of results between data sets
 - 6.4.1 Task difficulty
 - 6.4.2 Task-to-test fit
- 6.5 Summary

6.0 Introduction

This chapter reports the results of the datasets described in chapter five, the expert panel (section 6.1), the content comparison against external tests (section 6.2) and the IRT analysis of completed test data (section 6.3). Full results are discussed and presented in

tables. As well as reporting results within datasets, two examples are given of comparison of evidence across datasets (section 6.4).

The types of analysis reported here feed into the development of the theoretical model for continuous validation in chapter eight, and chapter nine draws directly on these results in its application of the model to the Five Star test.

6.1 Data set 1 : expert panel

6.1.1 Expert panel stage 1: preliminary orientation activities and skills identification

The familiarisation round described in 5.1.2 invited panelists to comment on working definitions of four skills, Listening, Speaking, Reading and Study skills. Their anonymous comments on these definitions were collated to produce revised definitions with alternative definitions for each skill, which were circulated again for panelists to select from. This process of eliciting a consensus for the skills definition was straightforward, with 'listening' producing the greatest range of possible definitions, but 'study skills' being reported anecdotally as presenting the greatest difficulty in defining, due to its overlap with other skills such as reading. The final version chosen by consensus was used as the definition of that skill in subsequent panel activities.

After the familiarisation round, the first panel activity asked panelists to identify whether or not each task tested particular skills. The group consensus produced ranges

from a high degree of agreement over most tasks and most skills to a complete lack of agreement over four tasks.

The panelists' judgements at stage 1 are tabulated in full in Table 13 below. Rows show each task in the test; columns show how many of the 12 panelists scored *yes*, *no*, *unsure* and *no answer* scores for each skill for each task.

Taking into account the likelihood of patterns of responses occurring by chance, the criterion set for panel consensus was that at least 10 of the 12 panelists must agree that a particular skill was or wasn't tested by a particular task. Thus only the score patterns in Table 13 were considered significant.

Table 13 **Basis for panel consensus on skills allocation**

For any task:	yes	no	unsure or no answer
	12	0	0
	0	12	0
	11	1	0
	11	0	1
	1	11	0
	0	11	1
	10	2	0
	10	1	1
	10	0	2
	2	10	0
	1	10	1
	0	10	2

In other words, where 3 (25%) or more of the panelists did not explicitly identify a skill as being tested by a task, it was considered as not indicating consensus. There is some evidence to suggest that one or two panelists in particular tended to indicate only a 'yes'

answer, i.e. using the default 'no answer' to indicate 'no'. This strengthens the case for regarding as non-significant patterns of response such as in Table 14.

Table 14 Non-significant consensus on skills allocation

For any task:	yes	no	unsure or no answer
	9 or fewer	0	3 or more
	0	9 or fewer	3 or more

On this basis, only scores of 10, 11 and 12 in the 'yes' column signify panel consensus and the skills so identified are listed in the last column of Table 15. Rows in Table 15 show each task in the test; columns show how many of the 12 panelists scored *yes*, *no*, *unsure* and *no answer* scores for each skill for each task. The final column shows the skills consensus based on the response patterns described above.

Table 15 Panel judgements on skills allocation for each task

	Listening				Speaking				Reading				Study skills				consensus of skills tested
Task	y	n	un	n/a	y	n	un	n/a	y	n	un	n/a	y	n	un	n/a	
1-4	11	0	0	1	12	0	0	0	1	8	1	2	0	8	2	2	listening, speaking
2-5	11	1	0	0	11	0	1	0	5	6	0	1	12	0	0	0	listening, speaking, study skills
3-6	11	0	0	1	12	0	0	0	4	6	0	2	0	8	0	4	listening, speaking
4-7	11	0	0	1	12	0	0	0	3	6	1	2	1	7	1	3	listening, speaking
5-8	6	3	0	3	9	1	1	1	11	0	1	0	3	5	1	3	reading
6-10	12	0	0	0	12	0	0	0	1	8	0	3	0	9	0	3	listening, speaking
7-11	12	0	0	0	10	1	1	0	6	6	0	0	11	0	1	0	listening, speaking, study skills
8-12	12	0	0	0	12	0	0	0	1	8	0	3	0	8	1	3	listening, speaking
9-13	12	0	0	0	11	0	0	1	0	10	0	2	3	6	0	3	listening, speaking
10-14	12	0	0	0	0	9	0	3	2	7	0	3	11	0	1	0	listening, study skills
11-15	12	0	0	0	5	2	3	2	9	0	0	3	5	4	1	2	listening
12-16	12	0	0	0	3	3	3	3	9	1	0	2	6	2	1	3	listening
13-17	4	5	0	3	6	1	2	3	9	1	2	0	3	6	0	3	no clear consensus
X-18	7	1	2	2	11	0	0	1	6	3	1	2	10	1	0	1	speaking, study skills
14-19	10	1	0	1	9	1	1	1	12	0	0	0	6	3	1	2	listening, reading
15-22	12	0	0	0	0	10	0	2	0	9	0	3	7	2	2	1	listening
16-23	11	0	0	1	12	0	0	0	0	9	0	3	4	4	1	3	listening, speaking
17-24	6	3	1	2	10	0	0	2	0	9	0	3	5	5	1	1	speaking
18-25	1	6	2	3	3	6	1	2	12	0	0	0	5	3	2	2	reading
19-26	1	7	1	3	5	5	0	2	12	0	0	0	3	7	0	2	reading
21-28	1	8	0	3	3	5	1	3	11	1	0	0	5	3	1	3	reading
22-29	12	0	0	0	12	0	0	0	0	9	0	3	2	6	1	3	listening, speaking
23-30	10	1	0	1	12	0	0	0	12	0	0	0	2	6	2	2	listening, speaking, reading
25-33	10	0	1	1	12	0	0	0	0	8	1	3	3	4	3	2	listening, speaking
26-34	4	4	1	3	5	3	1	3	11	0	1	0	5	4	0	3	reading
27-36	1	8	0	3	4	4	1	3	11	1	0	0	6	4	0	2	reading
28-47	1	8	0	3	4	4	1	3	11	1	0	0	6	3	1	2	reading
29-50	9	0	0	3	12	0	0	0	0	8	0	4	7	3	0	2	speaking
30-51	2	5	1	4	5	2	1	4	11	0	1	0	4	4	2	2	reading
31-53	12	0	0	0	12	0	0	0	1	8	0	3	3	4	2	3	listening, speaking
32-54	12	0	0	0	12	0	0	0	3	5	1	3	10	1	0	1	listening, speaking, study skills
33-55	12	0	0	0	12	0	0	0	2	6	1	3	3	4	2	3	listening, speaking
34-56	12	0	0	0	12	0	0	0	2	6	1	3	4	3	2	3	listening, speaking
35-57	2	7	0	3	11	1	0	0	0	9	0	3	8	3	1	0	speaking
36-58	11	0	0	1	11	0	0	1	0	9	0	3	0	9	0	3	listening, speaking
37-59	2	7	0	3	12	0	0	0	0	9	0	3	9	1	1	1	speaking
38-60	9	0	1	2	5	0	3	4	0	5	2	5	1	4	2	5	no clear consensus
39-61	11	0	0	1	12	0	0	0	0	10	0	2	0	10	0	2	listening, speaking
40-62	11	0	0	1	12	0	0	0	0	10	0	2	0	10	0	2	listening, speaking
41-63	2	6	2	2	7	3	1	1	12	0	0	0	7	2	1	2	reading
42-65	3	5	2	2	12	0	0	0	1	9	0	2	1	7	3	1	speaking

Table 15 (continued) Panel judgements on skills allocation for each task

	Listening				Speaking				Reading				Study skills				consensus of skills tested
43-68	10	0	0	2	11	0	1	0	12	0	0	0	2	7	1	2	listening, speaking, reading
44-69	10	0	0	2	11	0	0	1	12	0	0	0	2	7	1	2	listening, speaking, reading
45-71	5	5	0	2	11	1	0	0	6	4	1	1	7	0	4	1	speaking
46-72	12	0	0	0	12	0	0	0	0	10	0	2	5	4	1	2	listening, speaking
47-73	5	5	0	2	12	0	0	0	6	3	2	1	8	1	2	1	speaking
48-74	12	0	0	0	12	0	0	0	2	8	0	2	7	3	0	2	listening, speaking
49-75	5	5	0	2	5	4	2	1	9	2	0	1	12	0	0	0	study skills
50-76	6	4	0	2	9	1	2	0	10	1	0	1	12	0	0	0	reading, study skills
51-88	4	6	0	2	8	2	1	1	10	0	1	1	12	0	0	0	reading, study skills
52-89	5	4	1	2	8	1	2	1	12	0	0	0	6	2	2	2	reading
53-91	5	5	0	2	7	1	3	1	12	0	0	0	12	0	0	0	reading, study skills
54-92	7	3	1	1	12	0	0	0	7	4	0	1	5	3	3	1	speaking
55-94	4	5	0	3	7	2	1	2	11	0	1	0	5	3	2	2	reading
56-97	12	0	0	0	11	1	0	0	6	4	0	2	4	6	0	2	listening, speaking
57-98	4	5	2	1	6	2	3	1	8	3	1	0	4	5	1	2	no clear consensus
58-100	12	0	0	0	12	0	0	0	0	10	0	2	3	6	1	2	listening, speaking
59-101	12	0	0	0	12	0	0	0	0	10	0	2	4	5	1	2	listening, speaking
60-102	2	7	1	2	10	1	0	1	11	0	1	0	4	5	1	2	speaking, reading
61-103	10	1	0	1	12	0	0	0	11	0	0	1	5	4	1	2	listening, speaking, reading
62-104	5	5	1	1	9	1	1	1	11	1	0	0	7	1	2	2	reading
63-105	10	1	0	1	10	1	1	0	12	0	0	0	3	6	1	2	listening, speaking, reading
64-107	2	8	0	2	4	5	2	1	11	1	0	0	12	0	0	0	reading, study skills
65-108	2	6	2	2	4	4	2	2	12	0	0	0	11	0	0	1	reading, study skills
66-109	2	7	1	2	8	3	0	1	12	0	0	0	3	5	1	3	reading
67-110	8	0	0	4	7	0	1	4	1	4	1	6	1	1	3	7	no clear consensus
68-112	2	7	1	2	8	3	0	1	11	0	1	0	3	4	3	2	reading
69-113	8	3	0	1	11	1	0	0	12	0	0	0	6	3	1	2	speaking, reading
70-114	4	5	1	2	11	1	0	0	12	0	0	0	7	3	0	2	speaking, reading
71-115	3	5	0	4	8	1	1	2	11	0	0	1	4	2	2	4	reading
72-120	2	7	1	2	9	2	0	1	11	0	1	0	2	5	3	2	reading
73-123	3	6	1	2	6	4	1	1	12	0	0	0	4	4	2	2	reading
mdn	8.0	1.0	0.0	2.0	10.5	1.0	0.0	1.0	7.5	3.0	0.0	1.0	4.5	4.0	1.0	2.0	
Avg	7.3	2.8	0.4	1.5	9.0	1.4	0.6	0.9	6.4	3.8	0.3	1.4	5.0	3.9	1.1	2.0	
SD	4.1	2.9	0.6	1.2	3.3	2.1	0.9	1.2	4.9	3.8	0.6	1.4	3.4	2.6	1.0	1.2	

Rows show each task in the test. Columns show the number of panellists out of a total of 12 scoring yes, no, unsure and no answer scores for each skill for each card. The final column reports the consensus among panellists on the skills tested by each card, based on a minimum of 10 of the 12 panellists explicitly identifying a skill as being tested by a card.

Overall breakdown of tasks for the whole test

The panelists' judgements yielded a group consensus ranging from a high degree of agreement over most tasks and most skills to a complete lack of agreement over a few.

Overall,

- 29 tasks showed a consensus that they tested a single skill only - see (a) below;
- 31 tasks showed a consensus that they tested two skills - see (b) below;
- 8 tasks showed a consensus that they tested three skills - see (c) below; and
- 4 tasks showed no consensus about which skills they tested.

Breakdown by skill combinations

(a) Of the 29 tasks showing a consensus that they tested a single skill only:

17 were reading

8 were speaking

3 were listening

1 was study skills

(b) Of the 31 tasks showed a consensus that they tested two skills:

20 were listening + speaking

5 were reading + study skills

3 were speaking + reading

1 was listening + reading

1 was listening + study skills

1 was speaking + study skills

(c) Of the 8 tasks showed a consensus that they tested three skills:

5 were listening + speaking + reading

3 were listening + speaking + study skills

Table 16 identifies the frequency of each skill in these results, individually and in combination.

Table 16 Frequency of skill allocations across all tasks

	(a) alone	(b) with one other skill	(c) with two other skills	Total of skill mentions
Reading	17	9	5	31
Speaking	8	24	8	40
Listening	3	22	8	33
Study skills	1	7	3	11
Total of skill combinations	29	62 (=2 x 31)	24 (= 3 x 8)	115

The totals of 'skills mentions' scores in the last column of Table 16 show that the panelists considered speaking to be the most commonly tested skill overall, followed by listening and reading, with study skills last. However, the great majority of speaking skill mentions (32/40) were in combination with one or two other skills, whereas most reading tokens (17/31) were a single skill, and indeed reading was the most frequent skill to be tested alone. All the tasks considered to be testing three skills included speaking and listening among the three. The results suggest that in this test at least speaking and listening are most commonly found in combination with each other, and sometimes with other skills.

Table 17 identifies the frequency of particular combinations of skills, whether in pairs or in a combination of three skills (it therefore 'double counts' some combinations compared with Table 16):

Table 17 Frequency of skill combinations across all tasks

	alone	With reading	With speaking	With listening	With study skills	Total alone and in combination
Reading	17	-	8	6	5	36
Speaking	8	8	-	28	3	47
Listening	3	6	28	-	4	41
Study skills	1	5	3	4	-	13
Total in combination		19	39	38	12	

In the skills definition in the familiarisation round described above, study skills was the most problematic of the skill areas to define, and this conceptual difficulty is borne out by the fact that at the bottom of Table 15 study skills shows the highest mean scores of the four skills for both *unsure* and *no answer*.

Conclusions

The results of stage 1 show that in most cases the panel was able to achieve a high degree of consensus about the skills being tested, giving evidence of construct validity of Five Star as a multi-skill test that reflects common-sense perceptions of skill combinations in real life, and providing preliminary data for adapting the algorithm to reflect a better balance of skills tested by each natural route. This information is complemented by the stage 2 results below which show in addition at what level of language proficiency the panel consider each item to require for successful performance.

6.1.2 Expert panel stage 2 : percentage skills allocation and determination of proficiency level

The panelists' judgements at stage 2 of the percentage of skills contribution to each task are tabulated in full in Table 18. The rows show the tasks in the test; the columns show the percentage of each task that the panelists allocated to each skill. The mean scores of their allocations for each task are reported separately for the two sub-groups of six panelists each, and then together for both groups, all 12 panelists together.

Table 18 Panel percentage judgements of skills underlying each task

Task		Mean scores - group one				Mean scores - group two				Mean scores - both groups			
		List	Spea	Read	StSk	List	Spea	Read	StSk	List	Spea	Read	StSk
1-4	Names	50.3	51.2	2.0	2.0	56.2	44.5	2.0	2.0	53	48	2	2
2-5	Base numeracy	23.7	45.3	8.7	22.8	46.2	31.2	2.8	27.0	35	38	6	25
3-6	School/study 1	51.2	51.2	2.0	2.0	42.0	49.5	10.3	2.8	47	50	6	2
4-7	School/study 2	51.2	51.2	2.0	2.0	42.0	49.5	10.3	2.8	47	50	6	2
5-8	Basic reading	10.3	23.7	65.3	2.0	6.2	34.5	58.7	3.7	8	29	62	3
6-10	School/study 3	48.7	53.7	2.0	2.0	42.0	52.8	2.8	2.0	45	53	2	2
7-11	Inter numeracy	14.5	29.5	12.0	40.3	38.7	26.2	4.5	24.5	27	28	8	32
8-12	Family/rec/tion	45.3	54.5	2.0	2.0	47.0	56.2	2.0	2.0	46	55	2	2
9-13	Al Harbis	52.8	48.7	2.0	2.0	57.8	42.0	2.0	4.5	55	45	2	3
10-14	Adv. numeracy	61.2	2.0	7.8	31.2	62.0	2.0	2.8	37.0	62	2	5	34
11-15	Student reports	76.2	6.2	6.2	7.8	75.3	7.0	8.7	11.2	76	7	7	10
12-16	Paper clips	75.3	7.8	6.2	6.2	80.3	6.2	7.8	7.8	78	7	7	7
13-17	Rdng 3 - Jeddah	2.0	25.3	58.7	12.0	2.0	32.8	64.5	4.5	2	29	62	8
X-18	Student grades	7.0	58.7	10.3	19.5	12.0	47.0	16.2	27.8	10	53	13	24
14-19	Rdng 4 - grades	12.0	22.0	61.2	3.7	19.0	24.0	56.0	9.0	15	23	59	6
15-22	Shapes 1	83.7	2.0	2.0	15.3	65.3	8.7	12.0	16.2	75	5	7	16
16-23	Vehicles 1	52.0	50.3	2.0	2.8	53.7	41.2	2.0	7.8	53	46	2	5
17-24	Footballers	9.5	79.5	2.0	9.5	2.0	92.0	2.0	7.8	6	86	2	9
18-25	Ladder	2.8	11.2	72.0	12.8	2.0	7.0	83.7	12.0	2	9	78	12
19-26	Kettle	2.8	11.2	72.0	12.8	2.0	7.0	83.7	12.8	2	9	78	13
21-28	Signs	2.8	3.7	83.7	7.0	2.0	7.0	88.7	11.2	2	5	86	9
22-29	Fridge	51.2	42.0	2.0	7.0	47.8	52.0	2.0	6.2	50	47	2	7
23-30	Reading 2	11.2	27.0	47.8	16.2	11.2	34.5	49.5	3.7	11	31	49	10
25-33	Traffic lights	7.8	67.0	2.0	6.2	14.5	77.8	2.0	9.5	11	72	2	8
26-34	Traffic lights 2	2.0	29.5	64.5	7.8	3.7	27.8	64.5	11.2	3	29	65	10

Table 18 (continued) Panel percentage judgements of skills underlying each task

27-36	Signs 2	2.8	4.5	82.0	7.0	2.0	2.0	88.7	9.5	2	3	85	8
28-47	Signs 3	2.8	4.5	82.0	7.0	2.0	2.0	88.7	9.5	2	3	85	8
29-50	road signs	4.5	82.8	2.0	8.7	14.5	78.7	2.0	9.5	10	81	2	9
30-51	road signs 2	2.0	29.5	65.3	4.5	2.0	29.5	62.0	12.8	2	30	64	9
31-53	Training center	50.3	50.3	2.0	4.5	52.0	46.2	2.0	6.2	51	48	2	5
32-54	Population	44.5	44.5	2.8	11.2	48.7	42.0	2.0	10.3	47	43	2	11
33-55	Kuwait City	50.3	49.5	2.8	4.5	52.0	46.2	2.0	4.5	51	48	2	5
34-56	Nagorno K	49.5	48.7	2.8	5.3	50.3	44.5	3.7	6.2	50	47	3	6
35-57	Making tea	2.8	87.0	2.0	7.0	2.0	83.7	2.0	12.8	2	85	2	10
36-58	Speculation 1	33.7	65.3	2.0	2.0	45.3	58.7	2.0	2.8	40	62	2	2
37-59	Puncture repair	2.8	84.5	2.0	8.7	2.0	89.5	2.0	8.7	2	87	2	9
38-60	Singapore	49.5	49.5	2.0	4.5	54.0	46.0	2.0	2.0	52	48	2	3
39-61	Speculation 2	42.8	53.7	2.0	2.0	49.5	51.2	2.0	2.0	46	52	2	2
40-62	Speculation 3	40.3	54.5	2.0	2.0	42.8	57.8	2.0	2.0	42	56	2	2
41-63	Road accidents	2.0	21.2	69.5	7.8	8.7	37.0	45.3	12.8	5	29	57	10
42-65	Regional affairs	4.5	88.7	2.0	7.0	5.3	77.0	16.2	7.8	5	83	9	7
43-68	Newspaper 1	12.8	18.7	55.3	12.0	6.2	43.7	49.5	7.8	10	31	52	10
44-69	Newspaper 2	12.8	20.3	57.8	12.0	5.3	29.5	66.2	7.0	9	25	62	10
45-71	Instructions	5.3	60.3	8.7	22.0	5.3	68.7	9.5	22.0	5	65	9	22
46-72	Lebanon	49.5	47.8	2.0	2.8	51.2	51.2	2.0	2.0	50	50	2	2
47-73	Instructions 2	6.2	62.8	8.7	22.0	4.5	68.7	10.3	19.5	5	66	10	21
48-74	Lille	48.7	46.2	2.0	4.5	50.3	50.3	2.0	4.5	50	48	2	5
49-75	Saudia timetable	6.2	17.8	33.7	42.0	13.7	16.2	38.7	38.7	10	17	36	40
50-76	Weather charts	3.7	13.7	40.3	43.7	2.8	10.3	42.0	47.8	3	12	41	46
51-88	Riyadh weather	6.2	7.8	40.3	47.0	3.7	15.3	34.5	50.3	5	12	37	49
52-89	Climatic change	6.2	6.2	68.7	19.5	6.2	5.3	77.0	17.0	6	6	73	18
53-91	Child death	3.7	11.2	47.0	41.2	2.8	8.7	54.5	41.2	3	10	51	41
54-92	Travel	16.2	56.2	8.7	17.8	10.3	66.2	6.2	19.5	13	61	7	19
55-94	Heathrow	2.0	21.2	72.0	5.3	2.0	17.8	74.5	7.8	2	20	73	7
56-97	Tim Severin	47.0	46.0	2.0	8.0	51.2	51.2	2.0	2.0	49	49	2	5
57-98	Free money	17.0	21.2	52.8	7.0	6.2	4.5	78.7	13.7	12	13	66	11
58-100	United Nations	50.3	50.3	2.0	2.8	51.2	51.2	2.0	2.0	51	51	2	2
59-101	US Hitech	50.3	50.3	2.0	2.8	51.2	51.2	2.0	2.0	51	51	2	2
60-102	UNIDO	2.0	22.8	65.3	12.0	2.8	23.7	67.8	9.5	2	23	67	11
61-103	Comp. priorities	24.5	50.3	21.2	5.3	20.3	49.5	33.7	6.2	22	50	27	6
62-104	Karoshi	9.5	10.8	70.8	10.8	6.2	3.7	83.7	7.0	8	7	79	9
63-105	Karoshi 2	11.2	16.2	69.5	2.0	3.7	26.2	69.5	4.5	7	21	70	3
64-107	Prices	2.8	10.3	45.3	46.2	3.7	5.3	51.2	47.0	3	8	48	47
65-108	Production	2.8	10.3	45.3	46.2	2.8	5.3	50.3	46.2	3	8	48	46
66-109	Porsche	2.0	23.7	67.0	10.3	2.0	13.7	75.3	10.3	2	19	71	10
67-110	Book review	50.0	50.0	2.0	3.0	51.2	51.2	2.0	2.0	51	51	2	2
68-112	Bosnia	2.0	23.7	66.2	11.2	2.0	13.7	77.8	8.7	2	19	72	10
69-113	Conservation	4.5	37.0	54.5	7.8	13.7	44.5	41.2	8.7	9	41	48	8
70-114	SA railway	2.8	41.2	32.0	24.5	2.0	52.0	41.2	8.7	2	47	37	17
71-115	Honey bee	4.5	7.8	62.0	24.5	4.5	2.0	82.0	12.0	5	5	72	18
72-120	Conservation	2.0	22.8	66.2	11.2	2.0	12.0	78.7	10.3	2	17	72	11
73-123	The computer	2.8	16.2	71.2	11.2	2.0	12.0	78.7	10.3	2	14	75	11
	mean scores:	23	36	30	12	24	36	33	12	23	36	31	12

The two panel sub-groups scored some individual tasks with widely different percentage skills allocations; in Table 18, for example, group one allocated 24% of task 2 (card 5) to listening but group two allocated it 46%. However, the two sub-groups agreed very closely on the overall proportions allocated to each skill, with the mean percentage scores reported at the bottom of Table 18 being within 3% of each other. This consistency between the panel sub-groups provides evidence for the reliability of overall panel judgements.

The identification of skills tested by each task in stage 2 largely bears out the consensus established in stage 1. Exceptions are

- tasks 13-17 and 57-98, where no significant consensus emerged at stage 1, gained over 60% allocation to reading at stage 2 (compare the results for task 13-17 in Table 15 and Table 18)
- task 14-19 where reading and listening were identified in stage 1, but at stage 2 speaking gained a higher percentage allocation than listening
- task 23-30, where listening, speaking and reading were all identified at stage 1, but at stage 2 listening was given only 11% of the allocation
- task 25-33 similarly identified listening at stage 1 but again only allocated 11% at stage 2
- task 38-60, which failed to achieve a consensus at stage 1 but seemed clearly divided between listening and speaking at stage 2
- tasks 43-68 and 44-69 where study skills achieved the same percentage allocation as listening in stage 2, but was not identified as significant at stage 1

- task 63-105 identified listening at stage 1 but allocated it only 7% at stage 2
- task 67-110 had no significant consensus at stage 1 but divided the allocation equally between listening and speaking at stage 2

Possible reasons for this variation between stage 1 and stage 2 results are the different methods by which the skill judgements were made, stage 2 being based on a numerical allocation to skills rather than the 'yes/no' procedure at stage 1; stage 2 decisions being made on the basis of greater exposure to each individual task and to the test as a whole, so that panelists had by then a greater degree of familiarity; and some lack of consensus over the degree to which understanding task instructions in English contributed to overall performance on certain tasks.

Determination of proficiency levels

Table 19 shows in full the minimum proficiency levels identified as being necessary to perform each task. These are reported in the same pattern as the skill allocations: the mean for sub-group 1 of the panelists, the mean for sub-group 2, and then the mean for all 12 panelists together. The external rating scale used for these proficiency judgements, described in section 5.1.2 above, is shown in Table 11 (Carroll and West, 1989).

The final right hand column of Table 19 shows a measure of the variation between panelists' judgements of task levels. This is calculated from quartiles using the formula for the quartile coefficient of variation V_Q (Kendall and Buckland, 1982)

$$V_Q = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

The rationale for using this formula is that unlike the standard deviation it does not make assumptions about equal distance between the levels on the ESU scale.

Table 19 Panel ratings for proficiency levels required for each task

Task	Task name	Mean rating group 1	Mean rating group 2	Mean rating both groups	Quartile coefficient of variation
1-4	Names	1.0	1.5	1.3	0%
2-5	Base numeracy	1.5	2.2	1.8	33%
3-6	School/study 1	2.0	2.7	2.3	6%
4-7	School/study 2	2.4	3.3	2.9	12%
5-8	Basic reading	1.8	2.4	2.1	6%
6-10	School/study 3	3.2	3.9	3.5	14%
7-11	Inter numeracy	2.2	3.8	3.0	33%
8-12	Family/recreation	3.2	4.2	3.7	16%
9-13	Al Harbis	3.3	4.0	3.7	14%
10-14	Advanced numeracy	4.4	4.8	4.6	23%
11-15	Student reports	4.1	4.2	4.1	14%
12-16	Paper clips	4.2	4.1	4.1	14%
13-17	Reading 3 - Jeddah	2.5	3.2	2.8	20%
X-18	Student grades	3.3	4.3	3.8	17%
14-19	Reading 4 - grades	3.5	4.1	3.8	7%
15-22	Shapes 1	4.1	4.2	4.1	14%
16-23	Vehicles 1	3.9	4.3	4.1	14%
17-24	Footballers	3.8	4.8	4.3	14%
18-25	Ladder	3.8	4.7	4.3	25%
19-26	Kettle	4.0	4.5	4.3	13%
21-28	Signs	4.2	4.8	4.5	13%
22-29	Fridge	5.0	4.8	4.9	0%
23-30	Reading 2	2.5	2.6	2.5	20%
25-33	Traffic lights	2.8	3.8	3.3	33%
26-34	Traffic lights 2	3.2	4.1	3.6	14%
27-36	Signs 2	4.2	4.7	4.4	14%
28-47	Signs 3	4.5	4.4	4.5	11%
29-50	road signs	3.8	4.2	4.0	25%
30-51	road signs 2	4.6	4.7	4.6	17%
31-53	Training center	5.2	5.4	5.3	12%
32-54	Population	5.5	5.5	5.5	10%
33-55	Kuwait City	5.7	6.4	6.0	17%
34-56	Nagorno K	6.2	6.6	6.4	9%
35-57	Making tea	3.8	4.6	4.2	9%
36-58	Speculation 1	5.2	5.6	5.4	9%
37-59	Puncture repair	4.8	5.3	5.0	8%

Table 19 (continued) Panel ratings for proficiency levels required for each task

Task	Task name	Mean rating group 1	Mean rating group 2	Mean rating both groups	Quartile coefficient of variation
38-60	Singapore	5.2	5.3	5.2	13%
39-61	Speculation 2	4.2	5.8	5.0	20%
40-62	Speculation 3	4.5	5.4	5.0	14%
41-63	Road accidents	5.1	5.4	5.3	9%
42-65	Regional affairs	5.3	6.3	5.8	8%
43-68	Newspaper 1	5.2	5.3	5.3	9%
44-69	Newspaper 2	5.2	5.7	5.4	9%
45-71	Instructions	5.7	5.6	5.6	9%
46-72	Lebanon	6.6	6.6	6.6	8%
47-73	Instructions 2	6.2	5.7	5.9	9%
48-74	Lille	5.8	6.3	6.0	10%
49-75	Saudia timetable	6.0	5.0	5.5	17%
50-76	Weather charts	6.0	6.5	6.3	17%
51-88	Riyadh weather	5.2	5.3	5.3	12%
52-89	Climatic change	5.3	6.0	5.7	11%
53-91	Child death	6.0	6.2	6.1	17%
54-92	Travel	4.2	5.7	4.9	12%
55-94	Heathrow	5.2	5.7	5.4	14%
56-97	Tim Severin	6.4	6.9	6.7	13%
57-98	Free money	5.3	6.3	5.8	11%
58-100	United Nations	6.6	7.3	7.0	12%
59-101	US Hitech	6.8	7.3	7.0	8%
60-102	UNIDO	6.0	7.2	6.6	20%
61-103	Company priorities	6.2	7.3	6.8	16%
62-104	Karoshi	5.5	6.5	6.1	17%
63-105	Karoshi 2	5.4	6.5	5.9	15%
64-107	Prices	6.8	6.9	6.8	11%
65-108	Production	6.4	6.6	6.5	8%
66-109	Porsche	6.6	6.8	6.7	9%
67-110	Book review	6.6	7.3	7.0	13%
68-112	Bosnia	6.9	6.9	6.9	5%
69-113	Conservation	7.3	7.6	7.5	7%
70-114	SA railway	6.5	6.9	6.7	12%
71-115	Honey bee	6.0	6.8	6.4	9%
72-120	Conservation	7.0	6.8	6.9	2%
73-123	The computer	4.9	5.5	5.2	20%
	<i>average scores:</i>	4.76	5.27	5.02	13%

The range of average minimum proficiency level ratings is from 1.3 (task 1-4) to 7.5 (task 69-113), with a general increase in difficulty level over the numerical sequence of the tasks in the test. The overall 'grand mean' of proficiency level ratings is level 5, suggesting that the average level of difficulty might be broadly defined as 'intermediate'

with the range from elementary to advanced (7.5 would represent approximately the level of the Cambridge Proficiency in English examination).

Looking at the measure of agreement between panelists' ratings in the last column of Table 19, the values for the quartile coefficient of variation vary considerably. Tasks that the panelists appeared to have had more difficulty agreeing on were 2-5, 7-11, 10-14, 13-17, 18-25, 23-30, 25-33, 29-50, 39-61, 60-102 and 73-123.

Possible reasons for this are

1. that these tasks themselves are more 'multi-level' than others and can satisfactorily be performed at differing levels of proficiency; a slight tendency for harder tasks to show a smaller degree of variation could be interpreted as lending some support to this.
2. that the panel members are using different personal yardsticks of 'successful' and different interpretations of the level definitions.
3. the inherent variability in the large number of subjective judgements being made.
4. some of the tasks with the greatest coefficient of variation - e.g. 2-5, 7-11 and 10-14 - involve numeracy, and this suggests a wider disagreement between panelists over the importance of language proficiency alone in such tasks. Further investigation could reveal whether the extent of this variation is linked to the panel's uncertainty over the allocation of study skills, reported above.

Conclusions

The skills allocation at stage 2 offers a more refined panel consensus on which tasks test which skills, with the four tasks that achieved no consensus at stage 1 all presenting a

clear skills profile at stage 2. The percentage allocations to skills will provide specific data on which to plan the distribution of skills/tasks and refine the skills allocations in the scoring algorithm.

The panel's consensus for the difficulty level of each task will also assist in the routing and scoring of tasks in the algorithm. The distribution about a mean of level 5 on the external 9 band scale gives some validity to the overall range in difficulty level of the test as a test of general proficiency whose mid-point is at intermediate level. Where tasks which appear to present particular problems in achieving level consensus are retained in the same format, attention to the consistency of judgements made by different assessors on those tasks will clarify whether the main source of the variation is human subjectivity or inherent variability in the task itself.

During stage 2 of the exercise, panelists also examined the test tasks one by one and made a total of 280 individual open-ended comments on the test tasks and suggestions task by task for improving their content and presentation. A sample of their comments for the first 15 tasks is shown in Table 20.

Table 20 Sample of panellists' comments and suggestions: first 15 tasks

Card	Topic	Comments on tasks and suggestions for improvement
4	Names	<ul style="list-style-type: none"> • straightforward initial task • perhaps at this stage candidate could be asked to spell his name(s) which is commonly demanded in daily life [x3] / information boxes could be numbered • add Q "Can you spell that for me please?" • [improve it] possibly by use of written prompts - though this would shift the balance of skills (eg first name, family name) towards reading • [suggest] face on screen? • the prompts & interview notes might encourage candidates to spell their names out or to confirm that the examiner has achieved correct spelling on the screen (such a modification would clearly change the skill and level allocations) • more info - eg age, address? / pictures? • make deliberate error when typing name, give opportunity for correction/ it might be worth adding some prompts about where he/his family is from, how long he/they have lived (t)here, how it differs, which location he prefers, etc, to be used only as appropriate at the interlocutor's discretion. This would make it into more of a conversation-based card; it would change the nature of what is being assessed. /Why are there only two exits, Hesitant and Complete (there don't have to be three, but if other cards do, one might as well be consistent?) / What are these to measure? Isn't there is a mismatch between these rating labels and the explicit criterion of Comprehension/Pronunciation? / light bulb doesn't seem to work
5	Base numeracy	<ul style="list-style-type: none"> • pictures could be more exciting; the card is crowded • I think it needs to be made more explicit whether the ? suggestions are to be given in English or L1 • too rapid progression from low numbers to relatively high numbers / [suggest] pictures; simpler number sequence vertically • [suggest] numbers in context, eg telephone numbers, times of day • 1-storey house without chimney pots? • there are three sections to this test [card]. It would be nice if the user could 'zoom in' on the section being dealt with at any one time • I like the objectivity of the scoring system here. Is it in fact going to be hard for novice interlocutors to keep the score in mind as the task progresses? / I agree about contextualisation. Could the three sections be tied in together?
6	School/study 1	<ul style="list-style-type: none"> • EFL seems to be an obscure prompt • is this a reading and then a speaking exercise? It's not clear • for level 1 questions may need to be simplified further • the training procedure/operating manual will need to make explicit how the interlocutor uses the <? > icons - in front of the candidate? If so, is there a risk that the candidate will think it is a reading task, or will perceive that he will have an advantage if he can read the prompts himself? Or is this deliberate with the labels anyway?

Table 20 (continued) Sample of panellists' comments and suggestions: first 15 tasks

7	School/study 2	<ul style="list-style-type: none"> Is the examiner able to develop any of the points to make 'conversation' more natural are the prompts primarily for the examiner or candidate? is this a reading and then a speaking exercise? It's not clear how many years were you at school? → 'For how many years did you go to school?' / picture of a school? for level 1 questions may need to be simplified further would it be feasible to add one or two more non-factual questions, eg what do you remember most about school? Can you remember your first day? Who was your best teacher (which subject)? Which if appropriate can be encouraged to lead in to more open-ended answers
8	Basic reading	<ul style="list-style-type: none"> 3. 'wenizt' * there's a mistake ... Did we do it?! [?? Has this version become corrupted??] / The sentences aren't connected and without a context asking candidates to read aloud. Seems to test very little regarding pron/intonation other than graphic/phonetic relationships with isolated words. Candidates are asked to read out loud isolated words so why have sentences at all? / [Suggestion for improvement: large circle with] isolated words drawn from dialogue [on left on screen]; simple dialogue including target phonemes [on right of screen] the card seems to be asking for pron & word recognition skills. Presumably meaning is not called for don't like the reading out of context - more difficult especially for low levels - should be a picture with sentences relating to picture & not words read in isolation, but sentences the instructions allow the examiner to select 5 words for the student to read aloud - it is the potential variation between words like I, it, he, a and words like pharmacy, restaurant, puncture which makes it difficult here to allocate level. Should the instructions limit the examiner to inviting the student to read content words? / a pointing hand instead of a cursor would allow examiner to select words for reading aloud, if clicking on them made them bold it might help also level of lexis required doesn't seem to equate with level of task. Should more familiar, everyday lexis used? Qu 3 needs misprint removed! would having some sentences very short & simple be more natural than picking words out of longer sentences? How could basically the same task be contextualised? - put sentences into an office context ?
10	School/study 3	<ul style="list-style-type: none"> application doesn't seem to be an obvious prompt but again this depends if the prompt is supposed to provide support to the candidate 1 first question should be more specific / 2 How / 3 Should clarify the skills focus I personally find the question 'what areas/skills do you find most difficult/most easy?' is often quite generative. As this is the highest level of the three School/Study cards, You could also put in, 'What do you think is the best age to learn a foreign language? Why?', or 'How do you think the teaching of English in schools could be improved?' to generate more open-ended answers
11	Inter numeracy	<ul style="list-style-type: none"> presumably 24/6 = 4 is understood? [suggest] contextualisation / opportunities for pics as in other cards, it would be useful to be able to switch of the voice in emergencies (eg some interruption in the testing situation) one way to contextualise would be to put the numbers in short sentence contexts, for listening, reading and formulae
12	Family/recreation	<ul style="list-style-type: none"> visual prompts seem OK part II of the task could be to say, "Now you ask me some questions about (Family, Recreationsetc)" using the same prompts

Table 20 (continued) Sample of panellists' comments and suggestions: first 15 tasks

13	Al Harbis	<ul style="list-style-type: none"> the picture bears no relation to the content a very difficult listening exercise - task = more difficult than language level. Alternative - student asked to identify which of 2 brothers is shown in pic on screen / picture not related to tape ... memory test. screen illustration doesn't help but it could - ie two pictures, one of each brother with cues to lifestyles, hobbies etc the duality/contrast makes a nice performance criterion for marking. A possible high-level extension might be a few questions like "Which brother would you expect to see at ... and why?"
14	Advanced numeracy	<ul style="list-style-type: none"> could the end of the listening test be indicated by a beep or tone? difficult listening exercise - [suggest] speech normal speed! I think this is an excellent bit of contextualisation. Should the interlocutor control each sentence one at a time, to make it less memory task?
15	Student reports	<ul style="list-style-type: none"> it is not clear whether the candidate marks down the grades or says them are they allowed to make notes? Think they should be able to could the end of the listening test be indicated by a beep or tone? another nice semi-objective marking criterion with a handy key
16	Paper clips	<ul style="list-style-type: none"> are they allowed to make notes? Think they should be able to test could be bigger. More obvious means of 'ordering' (physically) / I like the 'office' context, and the marking key
17	Reading 3 - Jeddah	<ul style="list-style-type: none"> no real comprehension necessary/how is comprehension tested? / no comprehension required here, barking at print Could do with some authentic follow-up task this doesn't test reading comprehension. Tests only ability to decode and pronounce. To test comprehension, needs some questions clearly depends on your view of the value of reading aloud! One variation might be to omit the names of the cities in the text, and to alter the map and text to make cross-reference from the map to the text necessary to fill the gaps. It could still be a reading aloud task (or not!)
18	Student grades	<ul style="list-style-type: none"> the grades might be better presented in the form of a graph or bar chart; this might stretch candidates more title of exam could be given in the instruction to the examine comes the phrase "use this opportunity for interaction, ie getting the students to request further information" I'm not at all sure what this means, or how, if it happened, the examiner could provide any further info requested I would have thought a bit of contextualisation is needed as an intro, isn't it? And perhaps one or two examples? Who are these students? What are their scores? Why are they important? Eg English language scores for job applicants!
19	Reading 4 - grades	<ul style="list-style-type: none"> the comprehension questions are very basic compared to the level of the text; these could be more demanding, eg what was the difference in percentage score between highest and lowest? [suggest] chart with text? how would having the screen present the questions affect the interaction? Would it 'depersonalise' it, and/or make it easier for the interlocutor to assess objectively? There could be a separate table of True/False questions that pop up

6.1.3 Expert panel stage 3: scoring video tests and identifying interaction strategies

Results and analysis

For reasons described in section 5.1.2 above, the skill of 'interaction' used by the test developers as one of the six original sub-skills was omitted from the first two stages of the panel exercise, but an additional task was added to the video viewing activity in the third panel stage to try to address the issue of interaction. The test developer suggested an initial list of six possible interaction strategies for panelists to watch for.

Because test events varied greatly in length and candidates attempted different tasks and different numbers of tasks, it would not be meaningful to compare directly the absolute number of observed uses of each strategy by each candidate. It was therefore decided to seek a measure of the relative extent of use by asking panelists whether a particular strategy was used by each candidate in each task, to add up these observed tokens of each strategy and compare them as a proportion of total use for each candidate.

Each panelist watched each video and observed the presence or absence of each the six named categories of verbal interaction strategy, task by task, and in addition noted any other verbal strategies for each task. Finally, at the end of each video test, they were invited to note any other features or skills, other than interactional strategies and the core skills of listening, speaking, reading and study skills evaluated in previous rounds, which in their opinion influenced the process or outcome of the test. The *pro formas* used for recording and collating these data is shown in Table 21.

MEMORANDUM

TO

FROM Nic

DATE 12 November 1996

REF Five Star Test validation: instructions for stage 3

Thank you for your help in the second stage of this project.

The final stage involves viewing video recordings of some authentic test administrations to assign an overall level of proficiency to each learner, and to identify whether there are verbal interaction strategies, other than the core language skills, which affect the overall performance. No access to the computer is needed. We also welcome further suggestions for improving the test.

The tests last between 10 and 40 minutes, and each has a sheet attached which lists the tasks attempted. There are two video tapes: tape A lasts 2 hours and contains 5 tests, tape B lasts 3 hours and contains six tests. Please watch tape **B** before you watch tape **A**. You may find the sound quality poor in some cases. You can rewind and replay, but please try not to be too detailed.

Watch through each test and on the attached sheets:

1. for each task, identify the verbal interaction strategies used by the learner. Examples of some interaction strategies are given below *. If the strategy used is already listed, put a tick in that column (1 - 6); if it is not, write it in the "Other" column (7).
2. for the test as a whole, indicate to what extent the interaction strategies in general **contributed** to the overall test performance
3. indicate to what extent the interaction strategies **detracted** from the overall test performance
4. identify any **other features or skills**, other than interaction strategies and the core skills of Listening, Speaking, Reading and Study skills evaluated in previous rounds, which in your opinion influenced the outcome of the test. Examples might be *guessing* or *general knowledge*
5. assign an overall level proficiency level from the scale 1 - 9 overleaf
6. Finally, make any further suggestions for improving the test on the separate sheet.

* Examples of the different interaction strategies might be:

column 1 confirms understanding	"I understand", "yes/yeah", agrees, disagrees, laughs etc.
column 2 seeks confirmation	"Do you mean..?", repeats with question intonation, etc.
column 3 seeks clarification	"I don't understand", "I'm sorry", "Please repeat", etc.
column 4 indicates need for clarification	Fails to respond/extended silence, "errr...", ermmm...", etc.
column 5 confirms own previous turn	"Yes", "That's right", or some equivalent
column 6 re-forms own previous turn	"No, I meant..." rephrases previous statement, etc
column 7 other	identify here any other verbal interaction strategy you notice

The tapes are in my letter tray in the TESOL office. There are several copies of each tape; the ones with green marker pen round the label are the masters, and have slightly better sound quality.

For this test as a whole:

2. To what extent did these **interaction strategies** contribute to the satisfactory completion of the tasks? Please tick one box:

significantly	moderately	insignificantly	not at all
---------------	------------	-----------------	------------

3. To what extent did the **interaction strategies** detract from the satisfactory completion of the tasks? Please tick one box:

significantly	moderately	insignificantly	not at all
---------------	------------	-----------------	------------

4. Are there any **other features or skills**, other than interactional strategies and the core skills of Listening, Speaking, Reading and Study skills evaluated in previous rounds, which in your opinion **influenced** the process or outcome of this test? If so, please identify them:

--

5. What is the **overall language proficiency level**, from 1 - 9 on the scale attached, demonstrated by the learner in this test?

overall language proficiency level

Table 21 (continued) Panel stage 3 *pro formas*

Finally, a major aim of this review and validation exercise has been to suggest how the test can be improved. Among other possibilities are greater use of audio excerpts, such as dialogue for example; short video clips; using colour; better graphics and/or animation; more detailed pop-up marking keys; and a separate writing test. Please make any recommendations, general or specific, for improvements to existing cards or ideas for new cards, topics or tasks.

Do you have any suggestions for improving the **existing** test tasks or topics?

Do you have any suggestions for **new** tasks or topics?

Do you have any other general suggestions or or recommendations for improving the Five Star test?

This is the end of the validation exercise. Thank you for your help!

The analysis of interaction for each of the video tests is shown separately in Table 22.

These analyses show the data for each task on each test and for each strategy from all the panelists together, so that a score of '10' for example, for interaction strategy one ('confirms understanding') for task 1-4 means that 10 out of the 12 panelists considered that that candidate used that strategy while performing that task.

Table 22 Panel judgements of candidates' interaction strategies

Numbers show number of panelists noting use of each strategy in each task. Key to strategy codes:

S1 = confirms understanding

S4 = indicates need for clarification

S2 = seeks confirmation

S5 = confirms own previous turn

S3 = seeks clarification

S6 = re-forms own previous turn

Test A 1	S1	S2	S3	S4	S5	S6	T	other interaction strategies (some duplicates removed)
4 names	10	0	0	2	10	4	26	<ul style="list-style-type: none"> repeats question before answering agrees with interviewer suggestions apologises adds information
11 inter numeracy	10	7	2	4	2	6	31	<ul style="list-style-type: none"> excuses/explains performance (difficulty in hearing) /seeks explanation of symbols apologises for errors repeats number while pointing / responds to text, adds comment
13 Al Harbis	7	0	5	3	3	2	20	<ul style="list-style-type: none"> indicates dissatisfaction with own response apologising for poor performance
30 reading	6	1	0	0	1	0	8	<ul style="list-style-type: none">
32 writing	6	10	0	3	2	0	21	<ul style="list-style-type: none"> offers, apologises
TOTALS	39	18	7	12	18	12	106	
<i>Are there any other features or skills which in your opinion influenced the process or outcome of this test? If so, please identify them:</i>	<ul style="list-style-type: none"> writing copying, handwriting style memory worry about incorrect answers -> lots of apologising -> slight intrusion involvement in interactional chat during writing 							

Test A2	S1	S2	S3	S4	S5	S6	T	other interaction strategies (many of these are duplicates)
4 names	7	2	1	1	8	2	21	<ul style="list-style-type: none"> corrects error (on part of interlocutor when typing in name) agreeing; giving extra info
10 school/study 3	4	6	4	9	1	0	24	<ul style="list-style-type: none"> answers in Arabic [2] indicates he doesn't understand
11 inter numeracy	4	5	3	4	1	4	21	<ul style="list-style-type: none"> supplying unknown items in L1 guessing - going back repeating number
12 family/recreation	5	1	1	7	3	0	17	<ul style="list-style-type: none">

Table 22 (continued) Panel judgements of candidates' interaction strategies

13 Al Harbis	1	2	0	9	0	0	12	• expresses uncertainty
30 reading	2	1	1	0	0	1	5	• makes gesture to check • points to screen
32 writing	6	1	0	0	0	0	7	• indicates problem / spelling avoidance strategy
TOTALS	29	18	10	30	13	7	107	
<i>Are there any other features or skills which in your opinion influenced the process or outcome of this test? If so, please identify them:</i>								
<ul style="list-style-type: none"> • writing/copying English script • this candidate more than most (but by no means uniquely) gives many signals that he needs clarification but because this is a test he doesn't get clarification - in this sense his signals are not (in this testing situation) genuinely interactive - ie part of an exchange • speed of reactions - guessing - use of gestures / uncertainty about requirements • often only responding to interviewer's closed questions 								

Test A3	S1	S2	S3	S4	S5	S6	T	other interaction strategies (many of these are duplicates)
4 names	9	7	2	0	9	5	33	<ul style="list-style-type: none"> • freely giving information • greetings • asking questions, giving extra info • social interaction, eg explaining Arab name system
10 school/study 3	8	1	1	0	6	1	17	<ul style="list-style-type: none"> • gives extra information • expressing opinions • asking about test
15 student reports	11	10	5	0	2	3	31	<ul style="list-style-type: none"> • attempts to indicate wider knowledge • discussing other meanings of words • giving alternative explanations • speculates
16 paper clips	11	10	4	6	2	5	28	<ul style="list-style-type: none"> • expresses some answers tentatively to seek confirmation they are correct • offering alternatives • confirms exact task
22 shapes	11	4	0	2	3	1	21	<ul style="list-style-type: none"> • repeats target language • signals he can't do it • apologising for mistake / giving running commentary whilst using pointer on screen • repeating instructions, eye contact
23 vehicles 1	7	3	0	2	5	1	18	<ul style="list-style-type: none"> • requests pen • social interaction / talk about family
25 ladder	2	1	2	1	1	0	7	<ul style="list-style-type: none"> • discusses procedural matters, requests pencil & paper
28 signs	10	8	4	3	3	2	30	<ul style="list-style-type: none"> • indicates inability to answer • gives additional info/hypotheses
29 fridge	12	0	0	0	1	2	15	<ul style="list-style-type: none"> • paraphrasing • offers additional information
50 road signs	6	1	1	1	2	1	12	<ul style="list-style-type: none"> • uses gesture
55 Kuwait City	9	5	2	0	1	2	19	<ul style="list-style-type: none"> • paraphrasing, turn-taking • expressing degree of confidence • checking "I'm not sure but ..."
57 making tea	10	8	2	1	0	0	21	<ul style="list-style-type: none"> • elicits vocab from interlocutor • sequencing • checking "We can say ..." • seeks confirmation about task

Table 22 (continued) Panel judgements of candidates' interaction strategies

62 car ownership	9	1	0	0	7	6	23	<ul style="list-style-type: none"> • agreeing & expressing opinions • offering information
123 computer	8	1	1	3	0	0	13	<ul style="list-style-type: none"> • discussing difficulties with interviewer • eye contact for reassurance
TOTALS	123	60	24	19	42	29	297	
Are there any other features or skills which in your opinion influenced the process or outcome of this test? If so, please identify them:	<ul style="list-style-type: none"> • sequencing - interaction / very interactive with interviewer • memory • this student actually role-played card 15 (student reports) eg "I can give him excellent" • giving extra information / adding to answers / offering alternatives / asking questions to check / eye contact for reassurance, confirmation / high level of interaction skills - used positively for checking etc • explains personal problems with test, seeks confirmation, a great deal of eye contact 							

Test A4	S1	S2	S3	S4	S5	S6	T	other interaction strategies (many of these are duplicates)
4 names	9	2	4	4	4	0	23	<ul style="list-style-type: none"> • confirms tester's info • adding info
11 inter numeracy	8	4	0	2	1	7	22	<ul style="list-style-type: none"> • checking being understood • repeats for self-correction / hesitates for thinking time
13 Al Harbis	6	1	0	0	4	3	14	<ul style="list-style-type: none"> • strategic competence
15 student reports	11	2	0	0	1	1	15	<ul style="list-style-type: none"> • paraphrasing & strategic competence • translates to confirm understanding • reformulates examiner language • gives definitions
19 grades	6	2	0	4	2	1	15	<ul style="list-style-type: none"> • elicits vocabulary from interlocutor • strategic competence • (nods) • repeats examiners pronunciation
23 vehicles	3	4	6	3	1	0	17	<ul style="list-style-type: none"> • turn-taking • nods
24 footballers	2	2	1	2	0	3	10	<ul style="list-style-type: none"> • seeks confirmation of vocabulary • turn-taking • gestures
26 kettle	1	0	0	0	0	0	1	<ul style="list-style-type: none"> • turn-taking
28 signs	5	2	2	0	3	0	12	<ul style="list-style-type: none"> • turn-taking • adding information
29 fridge	8	3	2	5	1	0	19	<ul style="list-style-type: none"> • non-verbal features/ gestures
50 road signs	3	1	0	0	0	3	7	<ul style="list-style-type: none"> • seeks expansion by tester
55 Kuwait City	5	1	1	4	3	0	14	<ul style="list-style-type: none"> • gestures • indicates lack of understanding
62 car ownership	8	1	0	0	4	4	17	<ul style="list-style-type: none"> • seeks confirmation that interviewer understands • enquiry about interviewer knowledge • gives additional info, self-correction
TOTALS	75	25	16	24	24	26	186	
Are there any other features or skills which in your opinion influenced the process or outcome of this test? If so, please identify them:	<ul style="list-style-type: none"> • interaction • information transfer • here and in some other tests the examiner's willingness to simplify and even pidginise his own language will have contributed to interactive effectiveness. Additionally there were frequent uses of Arabic • good paralinguistic skills 							

Table 22 (continued) Panel judgements of candidates' interaction strategies

Test A5	S1	S2	S3	S4	S5	S6	T	other interaction strategies other interaction strategies (many of these are duplicates)
4 names	11	0	0	0	4	1	16	<ul style="list-style-type: none"> • confirms correct assumptions • gives extra clarifying information, turn-taking
10 school/study 3	8	0	0	0	5	5	18	<ul style="list-style-type: none"> • corrects false assumptions • gives extra clarifying information, turn-taking
47 signs 3	6	0	0	0	5	1	12	<ul style="list-style-type: none"> • gives extra clarifying information, turn-taking • explains choice
54 population	11	0	0	2	7	5	25	<ul style="list-style-type: none"> • using interviewer's words • para-linguistic features, turn-taking • [4] but in a very positive way, ie "I think I missed that" • expanding on topic - adding extra rephrases
71 instructions	10	0	0	0	6	9	25	<ul style="list-style-type: none"> • using interviewer's words • strategic competence • commenting on form of test • apologises, responds to nature of task / gives definition • uses discourse markers
74 Lille	10	1	4	0	4	6	25	<ul style="list-style-type: none"> • expands on previous turn • strategic competence • requests second listening / apologises • uses discourse markers
76 weather charts	10	8	6	0	2	2	28	<ul style="list-style-type: none"> • seeks clarif/tion of concept not lang • expands on interloc's explanation • questioning time • uses repetition • communicative competence • discusses own accuracy • apologises
95 travel	11	6	5	1	7	8	38	<ul style="list-style-type: none"> • reformulates interviewer's turn • negotiates task parameters • use interviewer words (convergence) • gives extra info, initiates conv. • expresses opinions • requests definition
101 US Hitech	11	3	0	3	5	9	31	<ul style="list-style-type: none"> • expansion of concepts + rephrasing interviewer's explanation • turn-taking: gestures • adding info • hesitates, personal response to task 'I missed it'
103 co. priorities	12	2	4	3	6	8	35	<ul style="list-style-type: none"> • paraphrasing • how can I say it? (filler) • extra linguistic features, extra info • "I mean ..." • gives definitions, paraphrases • uses discourse markers

Table 22 (continued) Panel judgements of candidates' interaction strategies

109 Porsche	6	2	1	2	2	2	15	<ul style="list-style-type: none"> • admits failure / evaluates own perf. • explaining • self correction
110 book review	6	2	1	4	0	2	15	<ul style="list-style-type: none"> • interviewer words (convergence) • strategic competence • "I was just going to ask you" • indicates lack of understanding
114 SA Railway	10	2	0	2	3	4	21	<ul style="list-style-type: none"> • interviewer words (convergence) • extra info, readiness to initiate • rephrases
TOTALS	122	26	21	17	56	62	304	
Are there any <i>other features or skills</i> which in your opinion <i>influenced</i> the process or outcome of this test? If so, please identify them:	<ul style="list-style-type: none"> • general knowledge of world, world events • good turn-taking & development if interaction • previous knowledge/experience • ability to 'think aloud' (used by this candidate in many of the tasks) • interaction, paraphrasing, interpreting, analysing • memory, general knowledge • confidence & desire to please examiner play large parts here • questioning - checking - giving additional info 							

Test B1	S1	S2	S3	S4	S5	S6	T	other interaction strategies (many of these are duplicates)
4 names	8	5	2	0	6	2	23	<ul style="list-style-type: none"> • gives additional personal information
10 school/study 3	10	0	0	0	2	2	14	<ul style="list-style-type: none"> • initiates • corrects interlocutor's assumption "My father is Saudi" • gives additional personal information
15 student reports	10	8	3	0	1	0	22	<ul style="list-style-type: none"> • seeks re-assurance "right now?" • gives definition
16 paper clips	8	6	3	4	1	3	25	<ul style="list-style-type: none"> • reassuring examiner • does task again • assessing own performance
47 signs 3	8	7	11	5	1	0	32	<ul style="list-style-type: none"> • nodding head - indicate confirmation • negotiates • 'thinking aloud' • "what do you want me to do?" • asks for clarification of task (reading out loud?)
53 training center	4	0	2	7	2	1	16	<ul style="list-style-type: none"> •
58 family size	7	1	5	2	4	5	24	<ul style="list-style-type: none"> • personalisation • initiates • signalling discomfort at excessively personal questions - "What's your name?" • takes direction
71 instructions	10	0	0	0	2	2	14	<ul style="list-style-type: none"> • (much more confident from here)
74 Lille	6	1	0	0	3	0	10	<ul style="list-style-type: none"> • spontaneous comment • offering additional (unrequested) information
76 weather charts	6	1	0	8	3	3	21	<ul style="list-style-type: none"> • nodding, confirming understanding • evaluates own performance • apologises, varies speed considerably

Table 22 (continued) Panel judgements of candidates' interaction strategies

92 travel	7	10	7	2	4	1	31	<ul style="list-style-type: none"> • questions content of test "how would you know if ...?" • negotiating, challenging • gives additional information / seeks clarification of task • makes suggestion
94 Heathrow	4	0	2	6	0	0	12	<ul style="list-style-type: none"> • hesitating -seeks encouragement to continue
123 the computer	3	0	0	1	0	0	4	<ul style="list-style-type: none"> • formulaic response trend of encounter "Thank you very much, it's been a pleasure"
TOTALS	91	39	35	35	29	19	248	
Are there any <i>other features or skills</i> which in your opinion <i>influenced</i> the process or outcome of this test? If so, please identify them:	<ul style="list-style-type: none"> • st tackled questions quite deeply at times and this took more time and delayed response • previous knowledge/experience • general knowledge • the 'manner' of the examiner, which was pleasant & encouraging at all times. This would give confidence to the candidate • memory. General knowledge? Info transfer (graphs) • paralinguistic signalling (as opposed to verbal) ie nodding, gross bodily movements etc • requests personal involvement with examiner 							

Test B2	S1	S2	S3	S4	S5	S6	T	other interaction strategies (many of these are duplicates)
4 names	9	0	0	0	10	1	20	<ul style="list-style-type: none"> • gives extra information: initiates
10 school/study 3	9	1	0	0	5	4	19	<ul style="list-style-type: none"> • [6] elaborates to facilitate understanding
15 student reports	10	1	0	1	1	2	15	<ul style="list-style-type: none"> • turn-taking
16 paper clips	11	5	6	2	5	5	34	<ul style="list-style-type: none"> • sequencing • What was that ...?
29 fridge	9	2	2	1	3	4	21	<ul style="list-style-type: none"> • querying interviewer • seeks reassurance "was it no. 6?" • using gestures
47 signs 3	8	4	4	8	4	2	30	<ul style="list-style-type: none"> • repeats information, queries task • uses humour "I'm a good driver!" • willingness to initiate • agreeing "Why not" • commenting on difficulty of task, 'too fast', 'it's impossible'
50 road signs	7	2	1	1	2	5	18	<ul style="list-style-type: none"> • extra info: ?avoidance strategy? • explaining ignorance
55 Kuwait City	7	2	0	0	4	4	17	<ul style="list-style-type: none"> • discussion
56 Nagorno K	5	6	2	5	3	0	21	<ul style="list-style-type: none"> • elicits vocab from interlocutor • turn-taking
59 puncture repair	4	0	0	0	5	3	12	<ul style="list-style-type: none"> • paraphrase to compensate lack/vocab • cohesion • indicates readiness • uses gestures / pauses for vocabulary
61 advancement & ed	1	0	1	0	2	1	5	•
65/67 regional affairs	6	1	4	2	5	4	22	<ul style="list-style-type: none"> • question & answer • points • checking intention of interlocutor
69 newspaper 2	3	0	0	3	2	1	9	•
73 instructions 2	3	2	0	1	1	1	8	•

Table 22 (continued) Panel judgements of candidates' interaction strategies

74 Lille	8	3	5	6	4	3	29	<ul style="list-style-type: none"> evaluating own performance repeats T's pronunciation / uses gesture / comments on task / apologises
88 Riyadh weather	6	6	3	4	2	3	24	<ul style="list-style-type: none"> study skills - interpretation of graph evaluating task "It's a game, eh? It's confusing"
89 climatic change	5	9	0	1	0	1	16	<ul style="list-style-type: none"> sequencing skills indicating readiness
123 the computer	1	3	0	1	0	2	7	<ul style="list-style-type: none"> questions / checking
TOTALS	112	47	28	36	58	46	327	
Are there any <i>other features or skills</i> which in your opinion <i>influenced</i> the process or outcome of this test? If so, please identify them:	<ul style="list-style-type: none"> political knowledge / previous knowledge/experience Pakistani teacher didn't help pronunciation / pronunciation unclear gestures - better communicator than Al Saawi. Better lang[uage] at beginning but poor reading, listening. Descriptors describe oral skills predominantly a lot of gesture used both to indicate understanding during listening and to enhance communicative effectiveness when speaking readiness to give extra information / to question & check 							

Test B3	S1	S2	S3	S4	S5	S6	T	other interaction strategies (many of these are duplicates)
4 names	9	0	0	0	5	3	17	•
10 school/study 3	8	1	0	0	5	3	17	<ul style="list-style-type: none"> correcting interloc's assumptions gives personal info, cataphora
11 inter numeracy	9	6	0	3	3	2	23	<ul style="list-style-type: none"> seeks confirmation of response
15 student reports	12	1	0	0	2	2	17	<ul style="list-style-type: none"> strategic competence?
16 paper clips	11	4	1	2	2	1	21	<ul style="list-style-type: none"> repetition rephrases suggestion
47 signs 3	8	5	1	0	1	1	16	<ul style="list-style-type: none"> paraphrase?
54 population	6	1	6	3	3	3	22	•
71 instructions	6	3	0	0	6	3	18	<ul style="list-style-type: none"> strategic competence requests thinking time 'just a minute'
74 Lille	7	0	6	4	1	5	23	<ul style="list-style-type: none"> avoidance strategy?
88 Riyadh weather	9	6	4	6	1	1	27	•
89 climatic change	5	3	3	0	1	1	13	<ul style="list-style-type: none"> study skills evaluating own performance
91 child death	12	6	5	2	3	0	28	<ul style="list-style-type: none"> study skills inviting comparison with earlier task
92 travel	10	1	0	0	5	4	20	<ul style="list-style-type: none"> asides "I lived in Jeddah..." initiates turn-taking negotiating values, agreeing social interaction, adds comment
94 Heathrow	5	3	1	2	2	2	15	<ul style="list-style-type: none"> repeats word / alters pronunciation
97 Tim Severin	5	2	2	5	2	4	23	•
98 free money	3	4	4	2	0	0	13	•
123 the computer	5	0	0	0	0	2	7	<ul style="list-style-type: none"> seeking evaluation from interv 'OK?'
TOTALS	133	46	33	29	42	37	320	
Are there any <i>other features or skills</i> which in your opinion <i>influenced</i> the process or outcome of this test?	<ul style="list-style-type: none"> had done similar test previously. Confident. Some linguistic ability in evidence speaking not v good but reading, listening good. ? not so much discussion here - better grade on test? seem to be fewer discussion tasks with card 98 the candidate requested paper & pencil - I think this helped produce result self-confidence - candidate actively seeks confirmation & clarification of tasks 							

Table 22 (continued) Panel judgements of candidates' interaction strategies

Test B4	S1	S2	S3	S4	S5	S6	T	other interaction strategies (many of these are duplicates)
4 names	10	4	4	6	3	0	27	•
10 school/study 3	5	5	8	7	1	0	26	• initiates conversation opener
11 inter numeracy	6	0	1	8	0	0	15	• subvocalising stimulus
13 Al Harbis	5	4	1	5	1	1	17	• ?avoidance?
14 adv numeracy	6	4	2	1	1	0	14	• study skills • a lot of paralinguistic interaction • study skills
30 reading	2	4	2	3	0	0	11	• repeats word / self-corrects pronunciation
32 writing	3	0	0	0	0	0	3	•
TOTALS	37	21	18	30	6	1	113	
Are there any <i>other features or skills</i> which in your opinion <i>influenced</i> the process or outcome of this test? Identify them:	<ul style="list-style-type: none"> • learner lacked confidence. Writing test seemed inappropriate • in card 14 the examiner's overt paralinguistic signals certainly prompted responses. I suspect the student would otherwise have missed a lot 							

Test B5	S1	S2	S3	S4	S5	S6	T	other interaction strategies (many of these are duplicates)
4 names	9	4	3	0	7	0	23	•
10 school/study 3	7	2	0	4	6	5	24	<ul style="list-style-type: none"> • [2] of vocabulary appropriateness • initiates: extra information, strategic competence: avoidance strategy • self-correction
11 inter numeracy	9	8	3	7	2	4	33	<ul style="list-style-type: none"> • asks for help by admitting lack of knowledge "not point ...?" • repeats number
13 Al Harbis	5	2	3	6	4	3	23	<ul style="list-style-type: none"> • seeks repetition • seeks evaluation • hesitates
14 adv numeracy	8	1	1	0	1	0	12	•
15 student reports	11	3	6	8	1	0	29	<ul style="list-style-type: none"> • avoidance strategy • points / uses gesture
19 grades	3	6	2	6	0	3	20	<ul style="list-style-type: none"> • starts pronouncing word and looks to interviewer for help • seeks correction pointing to screen
24 footballers	6	2	0	4	3	3	18	<ul style="list-style-type: none"> • strategic competence (attempts) • uses gestures / assesses own performance
26 kettle	5	3	1	2	0	1	12	•
TOTALS	63	31	19	37	24	19	198	
Are there any <i>other features or skills</i> which in your opinion <i>influenced</i> the process or outcome of this test?	<ul style="list-style-type: none"> • he doesn't seem to have a very good memory • memory • good communication skills - use of paralinguistic skills important 							

Table 22 (continued) Panel judgements of candidates' interaction strategies

Test B6	S1	S2	S3	S4	S5	S6	T	other interaction strategies (many of these are duplicates)
[nb particular problem here with sound quality]								
4 names	7	0	0	1	2	0	10	
11 inter numeracy	7	7	0	0	1	0	15	•
13 Al Harbis	5	1	0	1	1	1	9	•
15 student reports	10	0	0	4	1	1	16	• gives definition
17 reading - Jeddah	3	1	1	1	0	2	8	•
22 shapes	4	4	2	3	1	0	14	•
23 vehicles 1	4	0	8	1	3	1	17	• can you er ... • requests second listening at outset
24 footballers	1	0	0	1	0	1	3	•
26 kettle	2	0	0	1	1	1	5	• instructing examiner to wait
28 signs	6	7	1	0	0	0	14	•
54 population	6	0	2	5	3	1	17	• requests second listening
71 instructions	7	0	0	0	5	2	14	• rephrases / requests thinking time
74 Lille	6	1	0	2	0	0	9	•
75 Saudia timetable	11	2	4	0	2	0	19	• suggests a fault • queries exercise completeness • [3] by suggesting that something was missing from screen display • gives explanation for answer / gives opinion of task
91 Child death	9	3	0	1	2	1	16	•
92 travel	8	3	7	1	5	3	27	• reformulating question • joking (suggesting score of 6 on a scale of 1-5) & agreeing/ disagreeing / negotiating/ justifying & expressing opinions* • offers additional information / personal opinion
TOTALS	96	29	25	22	27	14	213	
<p><i>Are there any other features or skills which in your opinion influenced the process or outcome of this test? If so, please identify them:</i></p> <ul style="list-style-type: none"> • <i>previous knowledge/experience</i> • <i>introverted character, reluctant to participate in verbal interaction. Little eye contact</i> • <i>information transfer</i> • <i>* I think the personalisation element in card 92 contributed to the burst of uncharacteristic communicative behaviour. It is a very good test and perhaps saved this candidate's bacon</i> • <i>apparent reluctance for eye contact</i> • <i>very difficult to hear, the candidate tended to speak into his hands. Inaudibility at time made [overall proficiency level] difficult to judge</i> 								

Taking test A1 as an example, the candidate was considered by ten panelists to use strategy 1 (confirms understanding) in the first task, by ten panelists again to use the same strategy in the second task, by only seven panelists in the third task and 6 panelists in the fourth and fifth tasks. This strategy was the most prominent and accounted for

37% of all the strategy tokens noted by all the panelists observing his test, more than twice as much as the percentage score for strategy 2 (seeks confirmation) or strategy 5 (confirms own previous turn). The space for 'Other interaction strategies' and 'Other features or skills' allowed panelists to amplify or note significant observations; in this case, for example, that the candidate apologised frequently.

Aggregating the data in Table 22, the tokens for each strategy are totaled for each test in Table 23, and expressed as a percentage of the total strategy use by each candidate in Table 24.

Table 23 **Summary of interaction strategies by panel sub-groups**

Figures show the total number of strategy tokens noted by all panelists observing each test video

Interaction strategy		Group 1 combined scores						Group 2 combined scores						Both groups combined scores										
		S1	S2	S3	S4	S5	S6	Total	S1	S2	S3	S4	S5	S6	Total	S1	S2	S3	S4	S5	S6	Total		
Test A 1		19	11	3	5	12	9	59		20	7	4	7	6	3	47		39	18	7	12	18	12	106
Test A2		12	8	7	16	9	6	58		17	10	3	14	4	1	49		29	18	10	30	13	7	107
Test A3		67	35	14	12	30	17	175		56	25	10	7	12	12	122		123	60	24	19	42	29	297
Test A4		37	19	10	16	14	15	111		38	6	6	8	10	7	75		75	25	16	24	24	22	186
Test A5		62	17	13	8	35	34	169		60	9	8	9	21	28	135		122	26	21	17	56	62	304
Test B1		51	21	17	13	16	10	128		40	18	18	22	13	9	120		91	39	35	35	29	19	248
Test B2		64	27	16	15	35	28	185		48	20	12	21	23	18	142		112	47	28	36	58	46	327
Test B3		68	26	21	16	29	24	184		65	20	12	13	13	13	136		133	46	33	29	42	37	320
Test B4		21	10	11	14	6	1	63		16	11	7	16	0	0	50		37	21	18	30	6	1	113
Test B5		37	18	11	16	16	14	112		26	13	8	21	8	5	81		63	31	19	37	24	19	193
Test B6		50	17	15	14	13	8	117		46	12	10	8	14	6	96		96	29	25	22	27	14	213
Totals		488	209	138	145	215	166	1361		432	151	98	146	124	102	1053		920	360	236	291	339	268	2414

Table 24 **Percentage of use of interaction strategies, by panel sub-groups**

Percentages based on figures in Table 23

	Group 1 percentage scores for each strategy							Group 2 percentage scores							Both groups percentage scores						
Test	S1	S2	S3	S4	S5	S6	Total	S1	S2	S3	S4	S5	S6	Total	S1	S2	S3	S4	S5	S6	Total
A 1	32%	19%	5%	8%	20%	15%	100%	43%	15%	9%	15%	13%	6%	100	37%	17%	7%	11%	17%	11%	100%
A2	21%	14%	12%	28%	16%	10%	100%	35%	20%	6%	29%	8%	2%	100	27%	17%	9%	28%	12%	7%	100%
A3	38%	20%	8%	7%	17%	10%	100%	46%	20%	8%	6%	10%	10%	100	41%	20%	8%	6%	14%	10%	100%
A4	33%	17%	9%	14%	13%	14%	100%	51%	8%	8%	11%	13%	9%	100	40%	13%	9%	13%	13%	12%	100%
A5	37%	10%	8%	5%	21%	20%	100%	44%	7%	6%	7%	16%	21%	100	40%	9%	7%	6%	18%	20%	100%
B1	40%	16%	13%	10%	13%	8%	100%	33%	15%	15%	18%	11%	8%	100	37%	16%	14%	14%	12%	8%	100%
B2	35%	15%	9%	8%	19%	15%	100%	34%	14%	8%	15%	16%	13%	100	34%	14%	9%	11%	18%	14%	100%
B3	37%	14%	11%	9%	16%	13%	100%	48%	15%	9%	10%	10%	10%	100	42%	14%	10%	9%	13%	12%	100%
B4	33%	16%	17%	22%	10%	2%	100%	32%	22%	14%	32%	0%	0%	100	33%	19%	16%	27%	5%	1%	100%
B5	33%	16%	10%	14%	14%	13%	100%	32%	16%	10%	26%	10%	6%	100	33%	16%	10%	19%	12%	10%	100%
B6	43%	15%	13%	12%	11%	7%	100%	48%	13%	10%	8%	15%	6%	100	45%	14%	12%	10%	13%	7%	100%
Tot als	36%	15%	10%	11%	16%	12%	100%	41%	14%	9%	14%	12%	10%	100	38%	15%	10%	12%	14%	11%	100%

The pattern of strategy use discussed above for the first candidate A1 corresponds well with the overall pattern of strategy use, indicated in the total row of Table 24. Over all 11 test videos, strategy 1 accounted for 38% of the tokens; strategies 2 and 5 accounted for 15% and 14% respectively; and strategies 3, 4 and 6 for 12% or less.

Table 24 also shows as an indication of the reliability of the exercise that the two sub-groups of panelists came to roughly the same allocations, so for example, group 1 allocated 36% of its tokens to strategy 1 and 15% to strategy 2, while group 2 allocated 41% to strategy 1 and 14% to strategy 2.

Because the adaptive nature of the test gives candidates widely differing opportunities to employ these strategies, it is not possible to correlate candidates' proficiency levels against absolute numbers of strategy tokens. However, a comparison of the percentage breakdown of strategy tokens in Table 24 against the overall proficiency ratings in Table 26 confirms what one might expect: the stronger candidates (A3, A5 and B3) use strategy 4 ('indicates need for clarification') proportionately less than the other candidates, while the weaker ones (A2, B4 and B5) use it more. The weaker candidates use strategy 1 ('confirms understanding') and strategy 6 ('re-forms own previous turn') proportionately less than other candidates.

The panelists were also asked to make judgements on whether the interaction strategies they had observed contributed to or detracted from the candidate's performance in each test. This was to test the hypothesis that such interaction strategies have a beneficial

effect in communication. The summary of the judgements is shown in Table 25, by sub-group and as a total of all 12 panelists. The column headings 3, 2, 1 and 0 under 'contribute' and 'detract' in each case signify that the interaction strategies were judged to contribute (+) or detract (-) significantly (3); moderately (2); insignificantly (1); or not at all (0).

At the same time, the panelists made overall proficiency ratings for each candidate from 1-9 on the ESU scale (Table 11). The second and third columns from the right of Table 25 give the quartile coefficient of variation and mean of the proficiency ratings of all the panelists together; the coefficient of variation is calculated in the same way as reported in stage 2.2 above. The last column gives the candidate's actual Five Star test score, expressed in some cases as a range, e.g. 3-5, to reflect the Five Star test output being a profile rather than a single numerical score.

Overall, the panelists considered that the interaction strategies they noted contributed to rather than detracted from the satisfactory completion of the tasks. Of 103 panelists' judgements for a significant or moderate effect reported on the bottom row of Table 25, 85 (83%) were considered to have contributed to communication and 18 (17%) were thought to have detracted from it. There was a wide variation in the use of interaction strategies by different candidates and in the incidence of such strategies in different tasks. Since different candidates were presented with different tasks, a direct numerical comparison of their use of interaction strategies would be meaningless, but as an indication of the range, on the first task, which was common to all test administrations (task 1-4), between 16 and 33 incidences of interaction strategy were noted (discounting test B6 for reasons of sound quality). The numerical predominance of the first category

scored “confirms understanding” in almost every test indicates the significance of genuinely two-way interaction and its contribution to the successful negotiation of meaning, even in tests where the conversational initiative remains entirely or largely with the assessor.

Table 26 lists the individual and sub-group scores on the 1-9 ESU scale given by each panelist to each of the 11 video tests. The last two columns for each group gives the mean scores of those six panelists, and as for Table 25, the second and third columns from the right give the quartile coefficient of variation and the mean of the proficiency ratings of all the panelists together, and the last column gives the candidate's actual Five Star test score

Table 25 Summary of interaction contribution and proficiency ratings

Figures show total numbers of panelists' judgements that interaction strategies contributed to or detracted from performance

	Panel group 1						Panel group 2						Both panel groups						proficiency rating								
	strategies contribute			strategies detract			strategies contribute			strategies detract			strategies contribute			strategies detract			By expert panel (both groups)		By Five star test						
test	+3	+2	+1	0	-3	-2	-1	0	+3	+2	+1	0	-3	-2	-1	0	+3	+2	+1	0	-3	-2	-1	0	mean	coefficient of variation	
A1	2	3	1	0	0	1	3	2	0	4	2	0	0	0	3	3	2	6	4	0	0	1	6	5	2.8	20%	1
A2	0	1	2	3	1	1	3	1	0	2	3	1	0	0	2	3	0	2	5	5	1	1	5	4	1.6	33%	1
A3	4	2	0	0	0	0	2	4	3	2	1	0	0	0	2	4	8	4	0	0	0	0	4	8	5.4	9%	2-5
A4	2	4	0	0	0	0	5	1	1	2	3	0	0	1	0	5	3	7	2	0	0	1	5	6	4.1	14%	2-3
A5	4	1	1	0	0	0	1	5	1	3	2	0	0	1	0	5	8	3	1	0	0	0	1	11	7	14%	5
B1	0	4	1	1	0	2	2	2	4	2	0	0	0	0	0	6	0	10	1	1	0	2	3	7	5.4	9%	3-5
B2	0	4	1	0	0	2	3	0	0	6	0	0	0	0	1	5	2	6	3	0	0	3	4	4	4.5	11%	3-4
B3	3	2	0	0	0	0	1	4	2	2	2	0	0	1	1	4	3	7	1	0	0	0	2	9	5.6	16%	3-5
B4	1	2	3	0	0	3	0	3	0	5	1	0	0	0	1	5	2	2	6	2	0	3	1	7	2.5	23%	1
B5	1	3	2	0	0	1	2	3	1	0	3	2	0	0	1	4	2	6	4	0	0	2	2	8	2.6	22%	1
B6	0	1	2	3	1	2	3	0	1	3	2	0	0	1	0	5	0	2	5	4	2	2	3	4	3.9	21%	3-4
total	17	27	13	7	2	12	25	25	13	31	19	3	0	4	11	49	30	55	32	12	3	15	36	73			

Table 26 Video test proficiency ratings

Against the external rating scale in Table 11

	Panel group 1 by panelist (P1, P2 etc)							Panel group 2								Both panel groups		Five Star rating
test	P1	P2	P5	P10	P11	P12	mean	P3	P4	P6	P7	P8	P9	mean	mean	coefficient of variation		
A1	3	3	3	2	3	3	2.83	2	3	3	4	2	2	2.67	2.8	20%	1	
A2	2	1	2	1	1	2	1.50	2	1	2	2	2	1	1.67	1.6	33%	1	
A3	6	5	6	6	5	5	5.42	4	5.5	7.5	6	4	5	5.33	5.4	9%	2-5	
A4	5	3	5	5	4	4	4.33	5	2	5.5	4	4	3	3.92	4.1	14%	2-3	
A5	7	6	8	7	6	8	7.00	8	6.5	8.5	7	6	6	7.00	7	14%	5	
B1	6	7	6	6	6	4	5.75	5	5	6	6	3	5	5.00	5.4	9%	3-5	
B2	4	4	5		5	4.5	4.50	4	4.5	5	6	3	4	4.42	4.5	11%	3-4	
B3	5	6	7		7	4.5	5.90	6	6	7	5	4	4	5.33	5.6	16%	3-5	
B4	3	3	4	2	2	2	2.58	3	1.5	4	3	2	1	2.42	2.5	23%	1	
B5	2	2	3	4	2	3	2.58	2	2	4	4	2	2	2.67	2.6	22%	1	
B6	3	5	4	3	3	3	3.42	6	6	4.5	4	3	3	4.42	3.9	21%	3-4	
total	46	45	53	34	44	43		47	43	57	51	35	36					

A separate analysis of these data from the interaction strategies panel exercise by the test developer (reported in Pollard, 1997) concluded that much of the interaction took place in the instruction phase of certain tasks. The instructions are given in Arabic in less challenging tasks in order not to obscure the target skill being tested by lack of comprehension of instructions as an intervening variable, but higher-level tasks called on the interviewer to explain what needed to be done. Since these explanations were not the direct purpose of the test, but were intended to facilitate other tasks, they are arguably fully authentic, and the high incidence of interaction is therefore of extra interest (Lazaraton, 1992, uses a similar argument for the authenticity of the introductory phase of the oral proficiency interview, quoted in section 4.2 above). This finding influenced the subsequent development of the revised version of the test by splitting some tasks so that the explanation component has become in effect a separate task, and in some cases the candidate has to explain an Arabic instruction to the interviewer, reversing the conventional roles of the test participants.

As a note of caution, the appendices also show considerable variation between panelists in their perception and classification of interaction strategies, and any further studies into the nature of interaction in Five Star tests would need to address the standardisation of judgements. Yoshida-Morise (in Young and He, 1998: 228) also found unexpectedly low inter-rater reliability in classifying strategies particularly in respect of L1 (interlingual) strategies.

There is a clear relationship between proficiency level and the use of interaction strategies contributing to performance shown in Table 25, with the candidates with the highest proficiency ratings (A3, B1 and B3) having the most judgements that their use of strategies contributed significantly or moderately and the least judgements that they detracted significantly or moderately. This supports the study of Yoshida-Morise who found that six out of the eleven strategies she analysed showed significant differences according to candidate proficiency level (in Young and He, 1998: 225). What it does not justify is any inference about causality in the relationship between overall proficiency and the use of interaction strategies.

Table 26 shows a wide variation in the panelists' ratings of candidates' proficiency. The data support the view that most assessors are consistent in their deviation from the collective norm, in other words that some markers consistently over-mark and others consistently under-mark. It should be borne in mind that the panelists were using an external nine-point scale which was designed for general purposes rather than the specific three-point scale provided by the exit buttons on the test, and that they were making entirely subjective judgements in each case without reference to the pop-up exit criteria which in many cases offer more objective scoring guidelines. If consensus ratings from video tapes were to be used on a larger scale to validate individual raters' decisions, then a moderating exercise would be needed to standardise panelists' judgements.

There is also a close relationship apparent between the panelists' ratings of candidates' proficiency and the Five Star test scores. However, a numerical correlation coefficient

would be inappropriate here, as the Five Star scores are expressed as ranges while the proficiency ratings are scores on a single scale.

Conclusions

The apparent correlation between candidates' Five Star profiles and panelists' ratings of their proficiency contributes evidence towards test validity, but a formal concurrent validity study with a significant number of candidates against established external tests would be needed to confirm this.

The analysis of interaction patterns confirms that interaction is an important attribute of the Five Star test and that it can contribute to a candidate's performance. This provides substantial support for the construct validity of interaction as a feature of the test, but it does not help to define what interaction is or how it can be measured.

The analysis of interaction here cannot be interpreted in great detail, and indicates the need for further research both into the nature of the construct itself and its relationship with different task types, and candidates' personalities and proficiency levels. Poor sound quality was commented on by several panelists as making some of the interaction strategy judgements particularly difficult, and although the emphasis here was on verbal rather than non-verbal interaction strategies, the camera angle also served to obscure some of the finer interaction exemplars.

6.2 Data set 1: content comparisons

The content comparisons against three major established English language test systems (McCarthy, 1997; Graham, 1997; Kontoulis, 1997) were carried out by panel members who had specific knowledge of the criterion tests and were accredited examiners for those tests. They formed part of the critical review (Underhill, 1997) and are included verbatim at Appendix IV.

These are some of the points of commonality and contrast that emerged with the different criterion tests:

- a) all the criterion tests contain direct testing of speaking, like Five Star, and therefore an element of oral interaction. The UCLES main suite exams now have paired candidates and two examiners at each event; the Trinity College and IELTS tests, like Five Star, have a single candidate and a single interviewer.
- b) some of the tasks and topics are very similar across the four tests, for example, personal information and family relations, education, ambitions, profession or current study, etc. These are normal topics of conversation on which an individual can be expected to be able to talk without preparation.
- c) all the criterion tests cover the full range of proficiency from elementary to advanced, like Five Star. However, candidates for UCLES and Trinity College tests must be entered for pre-determined levels, whereas IELTS and Five Star are designed to accommodate candidates at any level at any test event.

- d) like the Trinity College exams, the Five Star consists only of tasks containing oral interaction, but it has a much wider range of task types. Like Trinity College, it is relatively short. The UCLES and IELTS tests are formal examinations with separate papers for testing reading, writing and listening skills.
- e) none of the criterion tests are adaptive or computer-based. Tasks in the criterion tests are delivered orally by the examiner or in print.
- f) although all the oral tests contain a sequence of tasks, the criterion tests, particularly at higher levels of proficiency, are more likely to elicit extended samples of spoken discourse from the candidate, through fewer tasks that take longer to complete. Five Star tasks can for the most part be satisfactorily completed by short utterances, although the scoring descriptors in some cases reward longer responses. Taken with the previous point, there is a tension in the test design between the desire to have more, shorter tasks that fit the adaptive model better, and to have fewer, longer tasks that allow extended interaction.
- g) all the criterion tests are aimed at a global market, and may be taken anywhere in the world. Five Star has a very specific local market, and is therefore strongly contextualised in its use of written and spoken Arabic and frequent reference to local issues and preoccupations. Topicality, in the geographical and temporal sense discussed in chapter four, raises very different problems for the criterion tests, such as how it is possible to have a communicative test that is universally relevant. Weir (1990), quoted in chapter two, considers it impossible to match test tasks with target language use on a global scale.
- h) the results of the Five Star and Trinity College test can be made known almost immediately, whereas the UCLES and IELTS tests may take several weeks to be marked.

Comment on the content comparisons

Content validation is essentially an intuitive process, as noted in chapter three, unless the target domain can be very clearly specified and the extent of 'closeness of fit' to a well-established and validated criterion test can then be used as a direct benchmark. As discussed in chapter two, this specificity is difficult to claim for language proficiency, in particular within a communicative paradigm that has not been fully and convincingly applied to language testing.

The external tests here were chosen as containing at least an element of direct assessment of spoken language and can to some extent be considered criterion tests. However, they are aimed at a global rather than a local market, and there are no other comparable tests with these features available for comparison that are specifically targeted at the same geographical market as the Five Star test. The contribution that these comparisons make may be more to construct than concurrent validity, following Kline's statement quoted in section 3.1 above, that concurrent validity "is only useful where good criterion tests exist. Where they do not concurrent validity studies are best regarded as aspects of construct validity" (Kline, 1993:19).

What content comparisons can therefore do is to highlight similarities and differences, without necessarily implying value judgements, which can be used to reflect back on the construct and the way in which the test matches the theory behind it. Two examples drawn from the comparisons above might be the desirability, in communicative terms,

of having a very specific geographical market; and the tension between the demands of adaptive test design on one hand and the possibility of extended interaction on the other.

6.3 IRT results and analysis

The data from 460 Five Star tests were analysed using QUEST (Adams and Khoo, 1996). The starting point for the analysis was a spreadsheet showing all scores on each task taken by each candidate. A small excerpt from this spreadsheet is shown at Appendix VI, displaying the results 30 candidates on 28 tasks. The total rows at the bottom of the appendix show the 'exit scores' for all 460 candidates. The mean time taken was 29.43 minutes, the mean number of tasks attempted was 9.7 and the median number of tasks attempted was 9.

The following sections report on the range of tables produced by QUEST to show estimates of case (candidate) ability, item (task) difficulty, error estimates and fit statistics. One of the unique features of IRT is the presentation of ability and difficulty on the same logit scale, so that a direct comparison can be made, for groups or individuals. Roughly speaking, if the difference between ability and difficulty estimates is positive, that item will be easy for that person. The more positive the difference, the easier the item and hence the greater the probability that the candidate will get that item right. A negative difference suggests an item that is going to be difficult for that person, and the more negative the difference, the greater the likelihood of failure on that item (Wright and Stone, 1979: 69).

The following section 6.3.1 looks at the summary statistics for item and case; the next section 6.3.2 considers the individual statistics for item estimates and section 6.3.3 looks at the individual statistics for case estimates. Section 6.3.4 looks at how QUEST

can display the distribution of all items and all cases on the same plot, and section 6.3.5 considers individual case (candidate) maps.

6.3.1 IRT outputs 1: summary statistics for items and cases

Table 27 shows the summary statistics for the item (task) estimates.

Table 27 Summary IRT statistics for the item (task) estimates

```

Item Estimates (Thresholds)                                16/ 6/99 17:51
all on all (N = 460 L = 73 Probability Level= .50)
-----

Summary of item Estimates
=====

Mean                                .00
SD                                  3.92
SD (adjusted)                       3.80
Reliability of estimate              .94


Fit Statistics
=====

Infit Mean Square                Outfit Mean Square

Mean      .91                    Mean      .99
SD         .28                    SD         .63


Infit t                            Outfit t

Mean      -.29                    Mean      .01
SD         1.23                    SD         1.17

0 items with zero scores
5 items with perfect scores

```

The mean of the case estimates is conventionally centred on 0. The different fit statistics are discussed in the following section.

Table 28 shows the summary statistics for the case (candidate) estimates.

Table 28 Summary IRT statistics for the case (candidate) estimates

```

Case Estimates
all on all (N = 460 L = 73 Probability Level= .50)
-----

Summary of case Estimates
=====

Mean                -1.32
SD                  3.44
SD (adjusted)       3.37
Reliability of estimate .96

Fit Statistics
=====

Infit Mean Square      Outfit Mean Square

Mean      .96          Mean      1.09
SD         .53          SD         1.79

Infit t              Outfit t

Mean      -.08         Mean      .26
SD         1.08         SD         1.13

0 cases with zero scores
0 cases with perfect score

```

The case estimates are measured on the same scale as the item estimates. A mean of case estimates of -1.32, compared with the mean of item estimates conventionally centred at 0, suggests that a typical item is 'difficult' for a typical candidate; in probabilistic terms, that there is a less than 50% chance of him getting that item right.

Fit statistics are discussed below.

6.3.2 IRT outputs 2: individual statistics for item estimates

The item estimates can be displayed item by item, with separate difficulty estimates for each threshold, and estimates of fit for the item as a whole. The full output of item estimates is shown in Table 29 below. For each item, the following values are shown.

SCORE shows that actual item score and MAXSCR the maximum possible score if every candidate who took the item reached the top score output. The THRESHOLD/S are the item difficulty estimates, in logits, for the exit scores for each task. The threshold is defined as the ability level required for a candidate to have a 50% chance of completing an item (McNamara, 1996: 291, after Masters, 1982). The bottom threshold has no difficulty estimate because it is the 'fail' score; once a candidate has been routed to a task, the bottom exit is the default option, and no separate difficulty threshold is calculated for this. The second and third exits for each task have difficulty estimates, conventionally centred on zero, with standard error estimates beneath them. An easier task will have a negative threshold value, a harder task a positive threshold value. In general, the earlier tasks in the Five Star test are easier and later ones are harder; thus in Table 29, the first ten tasks all have minus values for difficulty. If an item is discriminating properly, the difficulty value for the third threshold will be higher than for the second.

The last three columns in Table 29 show FIT statistics, the INFIT MEAN SQUARE, the OUTFIT MEANSQUARE and the standardised INFIT t value (the OUTFIT t value is also calculated, but omitted from this table; it is shown in the results in sections 6.3.1

and 6.3.3). The better an item fits the model, the closer the mean squares are to a value of one and the t-values to zero.

Table 29 QUEST output for individual item estimates

Item Estimates (Thresholds) In input Order 16/ 6/99 17:51
all on all (N = 460 L = 73 Probability Level= .50)

T A S K	SCORE MAXSCR		THRESHOLD/S			INFT	OUTFT	INFT
			1	2	3	MNSQ	MNSQ	t
1 4Names	673	920		-6.50 .38	-1.77 .28	1.03	1.69	.4
2 5Base numeracy	65	90		-10.26 .96	-6.91 .68	1.02	1.25	.1
3 6School/study 1	1	3		-8.28 1.41		.34	.31	-.9
4 7School/study 2	19	66		-7.31 .81	-2.02 1.52	.78	.67	-1.1
5 8Basic reading	1	3		-8.28 1.41		.34	.31	-.9
6 10School/study	586	708		-6.13 .44	-2.87 .36	.77	.45	-2.2
7 11Inter numerac	212	312		-6.91 .50	-3.24 .41	1.10	2.34	.9
8 12Family/recrea	20	116		-5.53 .69	-.40 1.78	1.03	.86	.2
9 13Al Harbis	148	394		-4.56 .34	-1.67 .41	.88	1.02	-1.2
10 14Advanced nume	42	146		-4.06 .50	-1.00 .73	1.11	1.03	.8
11 15Student repor	221	442		-1.94 .31	-.25 .33	1.37	1.37	3.3
12 16Paper clips	80	158		-1.50 .56	2.29 .58	1.31	1.64	1.9
13 17Reading 3 - J	85	142		-5.38 .59	-2.71 .53	.96	.87	-.2
14 19Reading 4 - g	32	134		-3.00 .56	-.89 .74	1.03	1.01	.2
15 22Shapes 1	75	180		-2.03 .47	1.50 .57	1.00	1.02	.1
16 23Vehicles 1	120	212		-3.47 .47	-.78 .42	.81	.78	-1.6
17 24Footballers	134	304		-3.25 .38	-.39 .41	.81	.81	-1.8
18 25Ladder	87	130		-2.84 .63	-.68 .54	1.12	1.12	.8
19 26Kettle	74	236		-2.47 .41	-1.02 .45	1.11	1.23	.8

*****Output Continues*****

Table 29 (continued) QUEST output for individual item estimates

Item Estimates (Thresholds) In input Order 16/ 6/99 17:51
all on all (N = 460 L = 73 Probability Level= .50)

ITEM NAME	SCORE MAXSCR		THRESHOLD/S			INFT	OUTFT	INFT
			1	2	3	MNSQ	MNSQ	t
20 27Writing	1	2		-9.86 1.41		1.00	1.00	.0
21 28Signs	69	186		-1.13 .44	-.59 .45	1.76	2.79	4.2
22 29Fridge	52	174		-.75 .44	1.81 .60	.73	.70	-2.1
23 30Reading 2	72	176		-6.81 .47	-3.92 .52	.84	.79	-1.2
24 31Writing	108	176		-11.57 1.31	-4.38 .55	1.03	1.00	.2
25 33Traffic light	10	40		-2.69 .97	-1.00 1.18	.71	.62	-.9
26 34Traffic light	3	13		-2.42 .70		1.16	.94	.5
27 36Signs 2	10	24		.06 1.22	1.67 1.26	1.79	2.11	1.9
28 47Signs 3	185	262		-.28 .41	.66 .39	1.41	4.25	2.9
29 50Road signs	79	164		-3.06 .56	1.15 .59	.89	.83	-.7
30 51Road signs 2	4	15		-1.46 .70		1.47	1.52	1.1
31 53Training cent	37	86		-.50 .72	1.92 .73	.81	.71	-.9
32 54Population	150	222		-.84 .50	1.09 .46	.81	.71	-1.5
33 55Kuwait City	28	160		-.03 .53	2.45 .95	1.03	1.10	.2
34 56Nagorno K.	16	66		.28 .75	3.30 1.36	.68	.63	-1.6
35 57Making tea	25	48		-3.94 1.66	2.17 1.37	.77	.63	-.3
36 58Speculation 1	64	86		-1.72 .91	.64 .65	.80	.75	-1.0
37 59Puncture repa	1	2		.44 1.43		.71	.71	-1.3
38 60Singapore	8	18		-.44 1.50	2.99 1.70	.65	.64	-.7
39 61Speculation 2	57	96		-1.63 .69	.41 .61	.75	.72	-1.4
40 62Speculation 3	79	150		-2.97 .53	.20 .54	.77	.78	-1.5

*****Output Continues*****

Table 29 (continued) QUEST output for individual item estimates

Item Estimates (Thresholds) In input Order

16/ 6/99 17:51

all on all (N = 460 L = 73 Probability Level= .50)

ITEM NAME	SCORE MAXSCR		THRESHOLD/S			INFT	OUTFT	INFT
			1	2	3	MNSQ	MNSQ	t
41 63Road accident	6	26		-.66 .50		1.13	1.76	.6
42 65Regional affa	33	76		-.94 .72	2.13 .79	.83	.82	-.7
43 68Newspaper 1	8	16		-.81 1.81	3.04 1.84	1.02	1.02	.2
44 69Newspaper 2	8	34		1.36 .43		.84	.73	-.7
45 71Instructions	246	344		-.66 .44	1.74 .35	1.08	2.36	.7
46 72Lebanon	11	32		.47 1.00	1.79 1.08	.72	.67	-.9
47 73Instructions	0	0	Item has perfect score					
48 74Lille	142	318		1.19 .34	3.69 .39	.80	.79	-2.1
49 75Saudia timeta	40	134		.50 .53	3.55 .81	.90	.88	-.7
50 76Weather chart	77	190		2.16 .44	5.26 .54	.93	.90	-.5
51 88Riyadh weathe	88	150		.78 .50	2.70 .47	.76	.95	-1.9
52 89Climatic chan	8	90		2.13 .88	2.68 .97	.71	.41	-.7
53 91Child death	6	39		2.89 .45		1.04	1.15	.2
54 92Travel	224	268		-1.47 .78	1.31 .42	.92	.71	-.6
55 94Heathrow	73	162		1.53 .50	4.36 .54	.83	.82	-1.3
56 97Tim Severin	18	30		2.38 1.16	4.31 1.04	.75	.70	-.7
57 98Free money	3	18		4.47 1.50	4.97 1.56	.72	.42	-.4
58 100United Natio	0	0	Item has perfect score					
59 101US Hitech	15	28		4.69 1.19	4.97 1.20	.88	.47	-.2
60 102UNIDO	2	6		5.49 .96		.48	.42	-1.2
61 103Company prio	6	8			4.68 .89	.74	.53	-.6

*****Output Continues*****

Table 29 (continued) QUEST output for individual item estimates

Item Estimates (Thresholds) In input Order									
all on all (N = 460 L = 73 Probability Level= .50)									
ITEM NAME	SCORE MAXSCR		THRESHOLD/S			INFT	OUTFT	INFT	
			1	2	3	MNSQ	MNSQ	t	
62 104Karoshi	5	18		3.88 1.34	5.42 1.58	1.16	1.38	.5	
63 105Karoshi 2	5	8		4.44 1.91	5.27 1.85	1.02	1.25	.2	
64 107Prices	1	4		5.95 1.35		.40	.28	-.9	
65 108Production	0	0	Item has perfect score						
66 109Porsche	7	26		5.59 1.22	6.72 1.32	.70	.79	-.7	
67 110Book review	7	18		5.84 1.41	7.60 1.59	.72	.65	-.5	
68 112Bosnia	0	0	Item has perfect score						
69 113Conservation	1	2			7.88 1.60	.36	.36	-.9	
70 114SA railway	14	26		3.09 1.41	6.51 1.28	.84	.82	-.3	
71 115Honey bee	0	0	Item has perfect score						
72 120Conservation	7	16		4.91 1.63	8.81 1.95	1.37	1.29	.9	
73 123The computer	76	158		-.22 .53	1.26 .56	.72	.60	-1.8	
Mean				.00		.91	.99	-.3	
SD				3.92		.28	.63	1.2	

Interpreting fit statistics

Broadly, the fit statistics show the extent to which individual items or cases fit the overall IRT model and the overall set of items in the test. Misfitting items may either be poor items, in the traditional sense of discriminating poorly, or they may be perfectly

good items in themselves, but be measuring something different from the rest of the items.

Authorities differ on how to set the confidence limits for interpreting fit statistics. Stansfield and Kenyon (1996: 132) suggest a range of -2 to +2 as limits on an item's goodness of fit as measured by the outfit mean square: 'all tasks with a standardized [mean square] outfit statistic above 2 were considered misfitting. Tasks with an outfit below -2 are considered overly consistent (not contributing unique information to the measurement)... Generally, less than 10% of the items should be misfitting before adequate fit is claimed'.

Wright and Masters (1982) suggest that the outfit mean square is more sensitive than the infit mean square to unexpected responses made by candidates for whom an item is much too easy or too difficult; Adams and Khoo (1996) agree that the infit statistic is more robust than the outfit, and add that 'it is closely related to item or case discrimination. Under most circumstances, infit and outfit values will be similar' (1996: 92).

Lumley uses the infit mean square, and sets a tighter limit, describing the acceptable upper limit for infit mean square at 1.3 before an item is to be considered misfitting (Lumley, 1993: 218). McNamara cites a range of 0.75 to 1.3 as acceptable for mean square statistics, and a range from +2 to -2 for values of t . He notes that t values may be inflated for large sample sizes, and considers 'the mean square statistic as safer as it is less sensitive to sample size' (McNamara, 1996: 173,181).

Although these limits appear to be set arbitrarily, McNamara also gives a formula for calculating the acceptable range for the mean square statistic as the mean plus or minus twice the standard deviation of the infit mean square statistic. On this basis, in the case of the Five Star item statistics, with an infit mean square mean of 0.91 and standard deviation of 0.28 (from Table 27 above), the acceptable range would be 0.35 to 1.47.

In the analysis in Table 29 there are five tasks with infit mean squares outside this range from 0.35 to 1.47. The five misfitting tasks are shown in Table 30.

Table 30 IRT statistics for misfitting tasks

Task	Score	Maxscore	Threshold 2	Threshold 3	Infit mean square	Infit <i>t</i>
3-6School/study 1	1	3	-8.28	-	0.34	-.9
5-8Basic Reading	1	3	-8.28	-	0.34	-.9
21-28Signs	69	186	-1.13	-.59	1.76	4.2
27-36Signs2	10	24	.06	1.67	1.79	1.9
28-47Signs3	185	262	-.28	.66	1.41	2.9

Five tasks misfitting represents 7% of the total of 73 tasks on the Five Star test, within the 10% limit suggested by Stansfield and Kenyon (1996).

Of these five tasks the first two, 3-6 and 5-8, have infit means squares of 0.34, which is just below the acceptable lower limit of fit of 0.35. Their *t* values are well within the acceptable range of -2 to +2. A likely explanation for their misfit lies in the fact that only very few candidates took these two tasks, and that the difficulty value in logits for threshold 2 was -8.28, in other words, extremely easy. There were no candidates reaching the higher score exits on these tasks. The problem with these two tasks may be

a lack of discrimination compounded by a very small sample on which to base the statistics.

The remaining three misfitting tasks, 21-28, 27-36 and 28-47, present a different challenge. There are respectable numbers of tests events reported for them, the t values lie outside or only just inside the limits of acceptability, and the difficulty values for all three tasks are in the middle of the range. Crucially, they are all the same type of task - identifying the appropriate text for signs in a particular context. They are discussed further in section 6.4.2 below.

6.3.3 IRT outputs 3: individual statistics for case estimates

A similar table to the item statistics is generated for the cases (candidates) whose test data forms the basis of the analysis. This is shown in full in Table 31.

Table 31 QUEST output for individual case (candidate) estimates

Case Estimates In input Order 16/ 6/99 17:51
all on all (N = 460 L = 73 Probability Level= .50)

NAME	SCORE	MAXSCR	ESTIMATE	ERROR	INFT	OUTFT	INFT	OUTFT
					MNSQ	MNSQ	t	t
1 001	10	21	-1.77	.52	.83	.74	-.32	-.41
2 002	5	14	-3.87	.75	1.25	1.05	.60	.32
3 003	7	14	-4.58	.74	1.21	1.07	.55	.33
4 004	5	14	-5.51	.77	.20	.22	-2.08	-1.58
5 005	14	18	2.04	.76	1.42	.88	.83	.49
6 006	6	10	-4.04	.87	1.07	.97	.31	.22
7 007	2	14	-8.12	1.09	1.07	.89	.33	.34
8 008	5	14	-5.70	.76	.54	.51	-1.00	-.77
9 009	7	25	-4.08	.60	.91	.73	-.06	-.09
10 010	9	20	-1.08	.63	1.02	1.28	.19	.64
11 011	17	22	2.56	.67	.80	.76	-.24	.44
12 012	6	10	-4.04	.87	1.07	.97	.31	.22
13 013	6	18	-4.57	.73	1.53	1.50	1.01	.82
14 014	6	14	-4.94	.75	.49	.53	-1.04	-.66
15 015	10	16	1.29	.67	.87	.69	-.03	.18
16 016	8	23	-3.73	.61	1.63	1.48	1.30	.91
17 017	12	21	-1.25	.50	1.17	1.02	.58	.21
18 018	23	32	1.35	.54	.76	.76	-.56	-.04
19 019	13	30	-2.07	.47	.81	.67	-.46	-.67
20 020	5	23	-4.96	.70	1.28	1.75	.67	.91
21 021	16	26	-.58	.49	.72	.84	-.83	-.16
22 022	8	14	-3.87	.73	.54	.54	-1.04	-.67
23 023	10	18	.75	.62	.75	.61	-.39	-.15
24 024	7	12	-1.53	.69	.99	.87	.14	-.01
25 025	10	23	-2.45	.53	1.02	.87	.19	-.12
26 026	5	16	-3.93	.73	.44	.40	-1.09	-.79
27 027	9	15	-.35	.65	1.86	2.15	1.54	1.45
28 028	6	14	-4.94	.75	.49	.53	-1.04	-.66
29 029	13	28	-1.78	.48	.96	.95	.00	.05
30 030	10	16	1.29	.67	.87	.69	-.03	.18
31 031	4	14	-6.94	.86	.61	.47	-.79	-.54
32 032	4	14	-6.83	.88	.90	.66	-.04	-.09
33 033	21	31	4.08	.58	2.15	1.42	2.06	1.01
34 034	5	12	-2.71	.72	1.27	1.36	.65	.71
35 035	18	29	.82	.49	.63	1.28	-1.04	.60
36 036	14	29	.42	.49	1.22	.90	.70	.03
37 037	9	14	-.35	.72	1.00	.91	.20	.15
38 038	6	14	-1.82	.70	1.82	1.39	1.61	.75
39 039	9	19	-1.62	.53	1.05	.83	.26	-.19
40 040	9	16	-1.34	.69	3.15	2.89	3.14	1.31
41 041	19	28	.53	.57	1.58	1.88	1.17	1.17
42 042	16	26	.21	.55	.75	.63	-.45	-.39
43 043	10	23	-2.23	.53	1.25	1.31	.70	.72
44 044	6	14	-5.46	.80	1.24	1.05	.59	.29
45 045	11	21	-1.48	.51	.81	.87	-.45	-.11
46 046	21	28	3.06	.66	1.39	1.08	.79	.66
47 047	9	17	-.55	.57	1.79	1.47	1.65	.88
48 048	10	22	-1.65	.53	.45	.48	-1.57	-1.06
49 049	4	14	-6.29	.79	.47	.42	-1.37	-.92
50 050	4	14	-4.21	.87	.79	.56	-.09	-.26
51 051	18	29	.80	.51	.77	.87	-.55	.01
52 052	6	14	-5.79	.81	1.84	1.56	1.51	.96
53 053	13	26	-.96	.49	.84	.68	-.40	-.60
54 054	11	20	-1.60	.59	1.51	1.36	1.28	.79
55 055	9	18	.25	.61	1.03	.68	.21	-.11
56 056	18	30	.34	.49	.47	.36	-1.49	-1.15

*****Output Continues*****

Table 31 (continued) QUEST output for individual case (candidate) estimates

Dataset analysis run 07 16th June 1999 'combine duplicates'

Case Estimates In input Order

16/ 6/99 17:51

all on all (N = 460 L = 73 Probability Level= .50)

NAME	SCORE	MAXSCR	ESTIMATE	ERROR	INFIT MNSQ	OUTFT MNSQ	INFT t	OUTFT t
57 057	14	21	1.59	.61	.89	.70	-.09	.07
58 058	4	14	-6.83	.88	.90	.66	-.04	-.09
59 059	11	26	-2.08	.52	.40	.38	-1.81	-1.39
60 060	6	20	-4.24	.66	.27	.24	-1.95	-1.37
61 061	5	14	-5.83	.79	1.62	1.59	1.19	.99
62 062	9	18	.34	.62	.28	.26	-2.13	-.87
63 063	7	12	-3.67	.79	1.32	1.57	.74	.93
64 064	16	26	.77	.52	.74	1.40	-.54	.71
65 065	19	26	2.06	.57	.47	.40	-1.40	-.24
66 066	16	24	1.26	.55	.78	.69	-.36	-.08
67 067	33	48	1.74	.39	.94	.78	-.12	-.03
68 068	6	10	-4.04	.87	1.30	1.20	.69	.50
69 069	6	25	-4.81	.68	1.00	1.05	.18	.32
70 070	15	28	-1.25	.48	.89	.82	-.25	-.26
71 071	17	20	3.60	.81	.22	.14	-1.62	.63
72 072	5	14	-6.11	.82	.55	.47	-.89	-.67
73 073	10	14	2.23	.78	1.04	.83	.25	.68
74 074	10	14	2.23	.78	.96	.76	.10	.64
75 075	7	13	.50	.82	1.81	2.41	1.30	1.29
76 076	4	12	-1.38	1.06	.37	.26	-.88	.06
77 077	6	14	-1.46	.80	.27	.26	-1.45	-1.17
78 078	12	14	4.15	.98	.34	.18	-1.05	1.76
79 079	10	14	2.23	.78	.58	.55	-.75	.52
80 080	17	23	5.18	.62	.72	.69	-.52	1.90
81 081	12	16	2.90	.76	.25	.19	-1.72	.46
82 082	10	14	2.23	.78	1.04	.83	.25	.68
83 083	14	18	4.18	.69	.49	.31	-1.01	1.36
84 084	9	13	1.45	.75	2.41	2.12	2.14	1.07
85 085	12	14	4.15	.98	.34	.18	-1.05	1.76
86 086	14	23	.10	.55	.72	.71	-.64	-.32
87 087	20	27	1.91	.56	.82	.63	-.29	-.05
88 088	0	7	-10.00	1.10	.84	.57	-.11	.02
89 089	13	26	-1.51	.50	.85	1.03	-.29	.22
90 090	6	14	-2.82	.68	1.20	1.18	.54	.49
91 091	11	16	2.36	.72	.34	.28	-1.60	.23
92 092	6	14	-2.82	.68	1.20	1.18	.54	.49
93 093	9	16	-.11	.66	.43	.37	-1.48	-.78
94 094	8	16	-.86	.73	.56	.49	-.69	-.56
95 095	15	29	-.62	.46	.56	.50	-1.61	-1.19
96 096	6	12	-2.22	.68	.99	.94	.13	.11
97 097	5	12	-2.71	.72	1.27	1.36	.65	.71
98 098	12	14	4.15	.98	.34	.18	-1.05	1.76
99 099	10	19	.16	.56	1.28	1.02	.78	.29
100 100	3	14	-7.80	1.01	.77	.43	-.10	-.22
101 101	6	19	-3.57	.64	1.02	1.14	.19	.43
102 102	10	18	.64	.62	.50	.40	-1.12	-.51
103 103	3	16	-7.62	.96	2.67	3.33	2.28	1.42
104 104	10	21	-1.77	.52	1.59	1.39	1.39	.87
105 105	6	19	-2.98	.62	.82	.82	-.28	-.07
106 106	9	14	1.90	.79	1.56	1.04	1.10	.67
107 107	3	14	-8.02	1.00	.79	.60	-.09	.05
108 108	8	21	-3.34	.60	2.01	1.92	1.95	1.48
109 109	3	14	-6.96	.87	.75	.58	-.41	-.32
110 110	2	14	-9.29	1.25	.11	.07	-1.21	-.13
111 111	17	20	6.89	.88	1.53	18.48	.87	5.36
112 112	6	11	-.81	.74	2.19	1.66	1.80	.99
113 113	12	16	2.55	.75	1.99	1.76	1.54	1.05
114 114	9	16	-.11	.66	.43	.37	-1.48	-.78
115 115	12	14	4.15	.98	.34	.18	-1.05	1.76
116 116	19	20	8.90	1.20	.54	.15	-.49	13.79
117 117	11	18	2.69	.75	1.89	1.47	1.35	.95

Table 31 (continued) QUEST output for individual case (candidate) estimates

Dataset analysis run 07 16th June 1999 'combine duplicates'

Case Estimates In input Order

16/ 6/99 17:51

all on all (N = 460 L = 73 Probability Level= .50)

NAME	SCORE	MAXSCR	ESTIMATE	ERROR	INFT	OUTFT	INFT	OUTFT
					MNSQ	MNSQ	t	t
118 118	8	16	.30	.63	1.22	.85	.59	.17
119 119	7	13	.50	.82	.28	.23	-1.57	-.68
120 120	10	14	2.23	.78	.58	.55	-.75	.52
121 121	11	19	.85	.58	1.16	.86	.50	.14
122 122	7	13	.32	.77	1.39	2.50	.78	1.36
123 123	23	28	6.43	.60	.88	.60	-.15	2.88
124 124	15	20	3.81	.67	.46	.32	-1.02	.91
125 125	12	18	3.19	.71	1.52	1.17	.93	1.02
126 126	7	14	-.86	.75	.35	.31	-1.27	-1.07
127 127	7	14	-.86	.75	.35	.31	-1.27	-1.07
128 128	10	16	1.86	.70	.93	.66	.00	.32
129 129	9	15	1.48	.74	.94	.94	.03	.42
130 130	4	9	-2.39	.98	1.11	.94	.37	.28
131 131	6	23	-4.11	.63	1.18	.92	.51	.08
132 132	19	22	3.40	.85	1.08	17.53	.35	2.55
133 133	1	14	-10.00	1.31	1.50	.88	.78	.88
134 134	5	18	-4.28	.73	1.22	.86	.56	.06
135 135	3	14	-8.02	1.00	.57	.38	-.50	-.23
136 136	5	17	-3.39	.66	1.04	1.10	.24	.36
137 137	12	14	4.15	.98	.34	.18	-1.05	1.76
138 138	8	21	-2.51	.54	.83	.86	-.30	-.09
139 139	8	13	1.16	.81	.75	.49	-.31	.01
140 140	2	14	-9.29	1.25	.11	.07	-1.21	-.13
141 141	4	20	-5.53	.79	1.21	.81	.55	.18
142 142	9	15	1.48	.74	.59	.44	-.81	-.05
143 143	9	14	1.65	.75	.56	.50	-.86	.21
144 144	9	19	-.21	.57	.87	.62	-.20	-.51
145 145	13	16	3.53	.83	.22	.15	-1.62	.90
146 146	7	19	-.86	.62	.58	.51	-.93	-.78
147 147	9	19	-.21	.57	.87	.62	-.20	-.51
148 148	7	14	-.86	.75	.35	.31	-1.27	-1.07
149 149	10	16	1.58	.67	1.36	1.07	.88	.53
150 150	9	15	1.21	.69	1.13	.86	.42	.30
151 151	8	14	.85	.71	1.51	.95	1.03	.36
152 152	21	34	1.82	.47	2.83	2.43	3.74	1.43
153 153	12	23	.61	.55	1.09	.85	.34	.03
154 154	13	18	1.93	.68	.48	.43	-1.10	.06
155 155	19	27	4.77	.57	.92	.58	-.08	1.28
156 156	10	18	.64	.61	1.10	.82	.36	.11
157 157	6	12	.67	.87	.53	.40	-.51	-.11
158 158	12	16	2.90	.76	.25	.19	-1.72	.46
159 159	7	12	1.35	.78	.87	.59	.03	.30
160 160	11	14	3.28	.89	1.02	.59	.26	1.16
161 161	8	22	-3.27	.62	1.86	2.14	1.69	1.43
162 162	27	36	4.11	.58	.99	3.17	.12	1.38
163 163	18	26	1.59	.54	1.87	1.65	1.71	.88
164 164	8	12	-1.00	.74	.76	.81	-.34	-.05
165 165	7	14	-.86	.75	.35	.31	-1.27	-1.07
166 166	6	13	-.31	.82	.73	.58	-.25	-.12
167 167	5	11	-2.05	.88	.91	.74	.04	-.06
168 168	9	15	1.48	.74	.59	.44	-.81	-.05
169 169	6	16	-3.43	.68	.44	.42	-1.22	-.90
170 170	12	20	.58	.62	.98	1.53	.12	.79
171 171	5	12	-.38	.94	.93	.61	.15	.12
172 172	9	14	-.35	.72	1.00	.91	.20	.15
173 173	15	21	1.96	.63	1.03	.92	.23	.38
174 174	7	19	-3.18	.60	.83	.82	-.30	-.19
175 175	20	34	.76	.44	.94	.88	-.08	-.02
176 176	17	31	.82	.46	.56	.52	-1.37	-.73

Table 31 (continued) QUEST output for individual case (candidate) estimates

Dataset analysis run 07 16th June 1999 'combine duplicates'

Case Estimates In input Order

16/ 6/99 17:51

all on all (N = 460 L = 73 Probability Level= .50)

NAME	SCORE	MAXSCR	ESTIMATE	ERROR	INFT	OUTFT	INFT	OUTFT
					MNSQ	MNSQ	t	t
177 177	15	18	3.38	.86	.88	.76	.05	.99
178 178	14	30	-1.95	.47	.56	.80	-1.29	-.29
179 179	11	23	.29	.55	.92	.79	-.09	-.08
180 180	11	16	2.36	.72	.71	.54	-.47	.45
181 181	12	23	.72	.53	.48	.48	-1.41	-.59
182 182	13	18	1.93	.68	.48	.43	-1.10	.06
183 183	11	14	3.28	.89	1.02	.59	.26	1.16
184 184	12	21	.44	.56	.19	.19	-3.06	-1.52
185 185	8	12	1.62	.88	1.84	1.17	1.37	.73
186 186	12	14	4.15	.98	.34	.18	-1.05	1.76
187 187	21	27	4.46	.67	2.99	21.48	2.79	2.67
188 188	19	26	.94	.60	2.81	2.46	2.73	1.53
189 189	17	25	1.52	.55	.53	.42	-1.20	-.47
190 190	16	26	-.23	.50	1.45	1.20	1.14	.52
191 191	7	12	-1.53	.69	.99	.87	.14	-.01
192 192	23	26	3.68	.79	.88	2.52	.00	1.30
193 193	4	10	-5.61	.91	1.96	1.73	1.38	1.04
194 194	9	14	1.65	.75	.52	.43	-1.01	.14
195 195	5	14	-5.51	.77	.20	.22	-2.08	-1.58
196 196	8	19	-.50	.59	.45	.43	-1.52	-1.03
197 197	14	21	1.49	.61	.70	.54	-.59	-.16
198 198	6	14	-4.94	.75	.49	.53	-1.04	-.66
199 199	14	20	2.07	.64	.49	.40	-1.18	-.01
200 200	5	14	-6.45	.82	.95	.86	.03	.03
201 201	13	27	-.88	.46	.60	.50	-1.44	-1.15
202 202	4	14	-6.14	.81	.72	.56	-.43	-.47
203 203	2	14	-8.12	1.09	1.07	.89	.33	.34
204 204	2	14	-8.12	1.09	1.07	.89	.33	.34
205 205	5	14	-6.45	.82	2.65	2.69	2.70	1.92
206 206	6	10	-4.04	.87	1.07	.97	.31	.22
207 207	15	24	1.02	.53	.75	.77	-.52	-.03
208 208	6	25	-4.45	.63	.97	.85	.10	.14
209 209	9	29	-3.61	.52	1.05	.88	.25	.02
210 210	4	14	-6.29	.79	1.18	1.51	.54	.89
211 211	7	20	-3.82	.63	.36	.35	-1.67	-1.18
212 212	14	25	.32	.48	.98	.82	.06	-.07
213 213	6	14	-2.82	.68	.85	.80	-.14	-.13
214 214	4	14	-6.29	.79	.47	.42	-1.37	-.92
215 215	6	20	-4.24	.66	.75	.54	-.37	-.54
216 216	3	18	-6.91	.91	1.21	1.50	.57	.83
217 217	26	40	1.39	.43	1.21	2.06	.72	1.41
218 218	20	27	1.11	.59	.82	.65	-.29	-.19
219 219	11	21	-1.58	.52	.70	.76	-.73	-.36
220 220	6	23	-4.92	.71	1.54	1.96	1.07	1.18
221 221	5	16	-3.93	.73	.44	.40	-1.09	-.79
222 222	5	20	-4.70	.70	.26	.21	-1.84	-1.21
223 223	10	26	-2.31	.51	.46	.41	-1.60	-1.27
224 224	7	16	-2.69	.64	1.26	1.11	.69	.38
225 225	18	26	.83	.58	1.34	1.44	.81	.75
226 226	7	14	-4.39	.73	.75	.69	-.40	-.33
227 227	11	19	.64	.59	.26	.24	-2.37	-1.05
228 228	14	20	2.07	.64	.49	.40	-1.18	-.01
229 229	8	12	-1.04	.72	.63	.66	-.68	-.32
230 230	2	14	-9.29	1.25	.11	.07	-1.21	-.13
231 231	14	29	-.82	.46	.35	.38	-2.83	-1.66
232 232	2	14	-8.12	1.09	1.07	.89	.33	.34
233 233	9	27	-3.45	.54	.80	.69	-.42	-.33
234 234	10	27	-2.58	.52	.57	.60	-1.20	-.67

Table 31 (continued) QUEST output for individual case (candidate) estimates

Dataset analysis run 07 16th June 1999 'combine duplicates'

Case Estimates In input Order

16/ 6/99 17:51

all on all (N = 460 L = 73 Probability Level= .50)

NAME	SCORE	MAXSCR	ESTIMATE	ERROR	INFT	OUTFT	INFT	OUTFT
					MNSQ	MNSQ	t	t
235 235	5	12	-2.71	.72	.87	.94	-.09	.12
236 236	21	26	2.79	.64	.56	.39	-.89	.10
237 237	8	12	1.62	.88	1.84	1.17	1.37	.73
238 238	13	21	1.22	.59	.67	.53	-.68	-.28
239 239	8	14	-.32	.71	.64	.52	-.63	-.49
240 240	15	19	4.48	.70	.77	.41	-.37	1.49
241 241	10	25	-2.82	.55	3.24	4.51	3.55	3.45
242 242	10	30	-2.82	.51	1.22	1.23	.66	.55
243 243	6	14	-4.94	.75	.66	.57	-.57	-.57
244 244	3	14	-7.19	.88	1.02	.77	.19	.02
245 245	7	14	-4.39	.73	.56	.53	-.91	-.67
246 246	5	14	-5.51	.77	.20	.22	-2.08	-1.58
247 247	13	24	.11	.50	1.14	1.07	.50	.30
248 248	5	14	-5.51	.77	.20	.22	-2.08	-1.58
249 249	7	16	-2.04	.67	1.77	1.37	1.61	.74
250 250	6	14	-5.21	.78	1.09	1.08	.34	.34
251 251	5	10	-4.80	.88	1.01	1.03	.22	.30
252 252	6	14	-5.46	.80	.88	.96	-.06	.16
253 253	16	24	1.16	.56	1.06	2.21	.28	1.22
254 254	6	12	-4.29	.78	1.45	1.84	.95	1.21
255 255	10	29	-3.31	.53	1.17	.97	.55	.15
256 256	5	20	-4.90	.71	1.23	1.17	.59	.47
257 257	17	24	.57	.58	.78	.63	-.44	-.35
258 258	13	29	-1.05	.45	.60	.46	-1.43	-1.32
259 259	21	26	2.33	.63	1.10	5.49	.36	1.93
260 260	10	16	1.29	.67	.86	.69	-.05	.17
261 261	5	14	-5.51	.77	.20	.22	-2.08	-1.58
262 262	8	12	-1.00	.74	1.55	1.32	1.10	.64
263 263	13	28	-1.19	.49	.77	.69	-.60	-.51
264 264	3	14	-7.19	.88	1.02	.77	.19	.02
265 265	19	23	5.32	.68	.62	.34	-.68	2.01
266 266	27	33	5.70	.61	.69	2.08	-.53	2.05
267 267	6	14	-3.01	.70	.92	.97	.00	.16
268 268	6	16	-3.43	.68	.68	.60	-.54	-.49
269 269	6	20	-4.24	.66	.45	.45	-1.21	-.76
270 270	5	14	-5.51	.77	1.07	1.01	.30	.23
271 271	17	28	-.25	.50	.91	.58	-.12	-.72
272 272	23	30	1.89	.59	.75	.98	-.51	.35
273 273	13	22	.04	.55	.49	.34	-1.50	-1.14
274 274	6	16	-3.03	.74	1.23	.95	.61	.22
275 275	4	14	-6.94	.86	1.25	2.13	.63	1.31
276 276	10	35	-3.02	.53	1.87	2.74	1.85	1.73
277 277	17	32	-.15	.47	2.13	1.88	2.43	1.38
278 278	6	12	.64	.90	1.14	.69	.43	.20
279 279	18	24	3.37	.72	1.33	1.39	.67	.96
280 280	17	28	-.25	.50	1.03	.70	.22	-.42
281 281	5	18	-4.68	.71	.69	.53	-.47	-.52
282 282	14	18	4.18	.69	1.03	.75	.23	1.51
283 283	10	21	-1.89	.54	1.68	1.44	1.71	.94
284 284	15	27	.87	.51	.83	.64	-.35	-.35
285 285	23	32	1.73	.51	.93	2.80	-.07	1.55
286 286	15	30	-.75	.46	1.21	.92	.71	.10
287 287	14	28	-1.48	.48	.87	.82	-.26	-.27
288 288	12	28	-2.06	.49	.49	.43	-1.51	-1.33
289 289	6	10	-4.04	.87	1.07	.97	.31	.22
290 290	13	28	-1.70	.47	1.05	.93	.26	.01
291 291	4	14	-7.16	.87	1.06	1.12	.28	.43
292 292	7	14	-4.84	.78	1.32	1.34	.72	.67
293 293	5	12	-2.71	.72	.87	.94	-.09	.12
294 294	22	31	1.72	.49	.87	.74	-.25	-.04
295 295	17	31	-.22	.48	1.26	1.19	.80	.52

Table 31 (continued) QUEST output for individual case (candidate) estimates

Dataset analysis run 07 16th June 1999 'combine duplicates'

Case Estimates In input Order

16/ 6/99 17:51

all on all (N = 460 L = 73 Probability Level= .50)

NAME	SCORE	MAXSCR	ESTIMATE	ERROR	INFT	OUTFT	INFT	OUTFT
					MNSQ	MNSQ	t	t
296 296	16	30	-1.38	.48	.97	.92	.02	.00
297 297	13	22	.82	.56	2.16	1.92	2.18	1.15
298 298	14	30	-1.81	.47	1.01	1.13	.14	.41
299 299	5	14	-6.27	.80	1.05	1.41	.25	.78
300 300	13	30	-1.61	.47	.96	.82	-.01	-.27
301 301	6	14	-4.94	.75	.49	.53	-1.04	-.66
302 302	6	18	-3.68	.65	.62	.82	-.71	-.07
303 303	10	18	-2.12	.59	1.27	1.27	.76	.65
304 304	14	21	1.49	.61	.59	.47	-.91	-.26
305 305	14	30	-1.95	.47	.56	.80	-1.29	-.29
306 306	10	18	.64	.62	.50	.40	-1.12	-.51
307 307	7	12	-1.77	.67	.71	.70	-.55	-.35
308 308	17	22	2.63	.66	.84	1.14	-.16	.68
309 309	6	23	-4.55	.68	1.21	1.36	.54	.68
310 310	3	14	-8.02	1.00	.57	.38	-.50	-.23
311 311	10	18	.64	.62	.50	.40	-1.12	-.51
312 312	5	10	-4.80	.88	1.01	1.03	.22	.30
313 313	15	30	-1.17	.48	1.00	1.27	.13	.65
314 314	9	14	-.35	.72	1.00	.91	.20	.15
315 315	10	24	-1.94	.51	.42	.36	-1.72	-1.49
316 316	7	13	.47	.72	.77	.60	-.30	-.32
317 317	10	22	-1.65	.53	.20	.22	-2.86	-2.07
318 318	5	14	-6.45	.82	.95	.86	.03	.03
319 319	9	16	-2.07	.64	.90	.88	-.09	-.04
320 320	5	14	-5.51	.77	.20	.22	-2.08	-1.58
321 321	20	24	2.76	.69	.84	.68	-.15	.42
322 322	16	27	-.14	.50	.75	.71	-.58	-.32
323 323	3	14	-7.80	1.01	.77	.43	-.10	-.22
324 324	17	33	-.38	.47	1.38	1.20	1.12	.51
325 325	20	31	-.05	.50	1.50	1.50	1.21	.89
326 326	6	14	-4.94	.75	.49	.53	-1.04	-.66
327 327	18	32	.79	.45	1.05	.81	.25	-.15
328 328	9	24	-2.85	.54	.82	.86	-.37	-.08
329 329	7	13	.32	.77	.81	.55	-.15	-.15
330 330	3	14	-6.96	.87	.75	.58	-.41	-.32
331 331	5	14	-5.51	.77	.20	.22	-2.08	-1.58
332 332	14	20	1.44	.62	1.85	1.38	1.46	.68
333 333	8	19	-.55	.59	1.17	.82	.53	-.12
334 334	14	23	2.01	.62	1.25	.85	.64	.34
335 335	9	20	-3.00	.59	1.52	1.40	1.20	.85
336 336	5	12	-.25	.99	.26	.20	-1.21	-.48
337 337	14	26	-1.19	.49	.94	.85	-.06	-.13
338 338	5	14	-6.27	.80	1.05	1.41	.25	.78
339 339	19	24	2.33	.63	.59	.51	-.85	.09
340 340	23	28	2.74	.64	1.00	2.09	.17	1.05
341 341	25	30	3.94	.63	.44	.24	-1.17	.46
342 342	3	14	-8.02	1.00	.57	.38	-.50	-.23
343 343	20	25	5.57	.64	.48	.24	-1.03	2.07
344 344	5	14	-3.54	.77	.73	.63	-.35	-.33
345 345	8	20	-2.70	.64	3.08	3.15	3.21	2.28
346 346	4	14	-6.29	.79	.94	.83	.01	-.05
347 347	16	25	.76	.50	.64	.64	-.99	-.36
348 348	9	18	.16	.64	.94	.70	.03	-.07
349 349	25	31	5.41	.60	2.32	5.09	2.40	2.11
350 350	14	21	-.49	.52	1.18	.99	.61	.16
351 351	22	29	1.80	.57	1.33	3.67	.87	1.80
352 352	10	16	.12	.68	.69	.43	-.64	-.59
353 353	6	10	-4.04	.87	1.07	.97	.31	.22

Table 31 (continued) QUEST output for individual case (candidate) estimates

Dataset analysis run 07 16th June 1999 'combine duplicates'

Case Estimates In input Order

16/ 6/99 17:51

all on all (N = 460 L = 73 Probability Level= .50)

NAME	SCORE	MAXSCR	ESTIMATE	ERROR	INFIT MNSQ	OUTFT MNSQ	INFT t	OUTFT t
354 354	4	14	-6.83	.88	.90	.66	-.04	-.09
355 355	5	20	-4.88	.72	1.25	1.16	.61	.47
356 356	27	36	1.85	.52	2.13	3.74	2.53	2.00
357 357	1	7	-10.00	1.10	.84	.56	-.11	.02
358 358	7	14	-4.39	.73	.68	.62	-.59	-.48
359 359	19	29	.26	.49	.38	.36	-2.00	-1.37
360 360	5	16	-3.93	.73	.44	.40	-1.09	-.79
361 361	17	28	-.74	.49	.77	.67	-.65	-.51
362 362	15	27	-1.33	.45	1.26	1.22	.88	.64
363 363	6	16	-3.43	.68	.68	.60	-.54	-.49
364 364	9	24	-2.21	.53	.43	.40	-1.62	-1.26
365 365	5	18	-4.68	.71	.69	.53	-.47	-.52
366 366	17	30	-1.24	.46	1.16	1.26	.57	.68
367 367	22	30	1.56	.57	.77	.84	-.47	.14
368 368	5	10	-4.80	.88	1.01	1.03	.22	.30
369 369	4	14	-6.94	.86	.61	.47	-.79	-.54
370 370	4	14	-6.29	.79	.47	.42	-1.37	-.92
371 371	14	25	-1.58	.49	.81	.83	-.45	-.24
372 372	5	14	-5.51	.77	.20	.22	-2.08	-1.58
373 373	20	30	.69	.53	.95	.75	.03	-.16
374 374	8	23	-3.14	.56	.60	.52	-1.02	-.72
375 375	4	14	-7.16	.87	1.06	1.12	.28	.43
376 376	5	14	-5.51	.77	1.14	1.19	.42	.50
377 377	8	27	-3.82	.55	1.42	1.57	1.05	1.08
378 378	17	26	.88	.55	1.46	1.36	1.09	.67
379 379	25	27	7.70	.89	.30	.10	-1.02	5.97
380 380	15	22	1.77	.60	.57	.50	-1.01	-.11
381 381	8	20	-3.36	.61	1.60	1.56	1.30	1.03
382 382	7	16	-2.04	.67	1.77	1.37	1.61	.74
383 383	6	19	-2.98	.62	.82	.82	-.28	-.07
384 384	15	23	1.20	.57	.50	.41	-1.29	-.57
385 385	4	20	-5.98	.80	1.32	.95	.78	.41
386 386	10	14	2.23	.78	.58	.55	-.75	.52
387 387	18	31	-.07	.46	1.01	1.96	.14	1.59
388 388	2	18	-6.93	.91	1.60	1.73	1.21	.93
389 389	8	19	-2.98	.58	.69	.65	-.72	-.61
390 390	8	12	-1.04	.72	.63	.66	-.68	-.32
391 391	15	20	2.51	.68	.39	.33	-1.36	.14
392 392	8	14	-1.07	.73	1.34	1.73	.78	1.13
393 393	5	14	-6.45	.82	.95	.86	.03	.03
394 394	6	16	-3.03	.74	1.23	.95	.61	.22
395 395	8	27	-3.98	.66	1.92	4.77	1.57	2.42
396 396	6	18	-3.79	.68	.96	.76	.08	-.16
397 397	14	26	-1.21	.48	.99	.89	.07	-.07
398 398	14	28	-1.05	.47	1.12	.95	.45	.04
399 399	2	18	-7.86	1.06	1.33	3.32	.66	1.46
400 400	12	26	-.93	.49	1.17	.83	.58	.04
401 401	17	24	1.62	.57	.52	.46	-1.22	-.29
402 402	21	26	2.79	.64	.35	.26	-1.60	-.05
403 403	9	22	-1.94	.55	.29	.30	-2.18	-1.62
404 404	12	28	-2.05	.51	.85	1.44	-.26	.92
405 405	19	26	2.06	.57	.51	.44	-1.24	-.19
406 406	16	18	4.20	.96	.33	.15	-1.07	1.33
407 407	25	31	3.80	.61	1.00	.86	.15	.71
408 408	12	26	-1.14	.54	1.52	1.24	1.30	.61
409 409	10	18	.62	.62	1.40	1.00	.90	.32
410 410	6	16	-3.43	.68	.68	.60	-.54	-.49
411 411	16	29	.33	.47	1.00	.99	.11	.16
412 412	8	14	-.08	.68	.56	.54	-1.00	-.40
413 413	16	24	-.25	.53	1.02	.70	.17	-.40
414 414	8	19	-.55	.59	1.16	.82	.49	-.12
415 415	15	26	-.48	.49	.81	.55	-.46	-.89

Table 31 (continued) QUEST output for individual case (candidate) estimates

Dataset analysis run 07 16th June 1999 'combine duplicates'

Case Estimates In input Order
all on all (N = 460 L = 73 Probability Level= .50) 16/ 6/99 17:51

NAME	SCORE	MAXSCR	ESTIMATE	ERROR	INFIT	OUTFT	INFIT	OUTFT
					MNSQ	MNSQ	t	t
416 416	12	31	-2.51	.52	2.97	2.83	3.45	1.84
417 417	2	14	-9.29	1.25	.11	.07	-1.21	-.13
418 418	14	23	.48	.53	.42	.33	-1.86	-1.14
419 419	20	28	5.11	.55	1.04	.48	.22	1.53
420 420	9	30	-3.77	.56	1.42	2.94	1.05	1.97
421 421	22	26	3.23	.70	.31	.20	-1.55	.14
422 422	12	22	-.36	.58	1.61	2.19	1.35	1.63
423 423	11	20	-.25	.56	1.29	1.07	.79	.31
424 424	18	28	.35	.54	2.05	1.62	1.91	1.00
425 425	7	14	-4.39	.73	.68	.62	-.59	-.48
426 426	9	20	-2.30	.60	2.02	1.85	2.10	1.38
427 427	14	22	1.14	.55	.48	.37	-1.22	-.56
428 428	10	27	-2.96	.54	1.56	2.87	1.30	2.29
429 429	3	14	-7.80	1.01	.77	.43	-.10	-.22
430 430	5	14	-5.51	.77	.20	.22	-2.08	-1.58
431 431	8	23	-2.44	.57	1.12	.98	.41	.16
432 432	22	33	2.05	.49	1.87	16.43	2.00	3.93
433 433	6	10	-3.94	.87	1.44	1.31	.86	.63
434 434	15	24	-.52	.51	.76	.54	-.60	-.89
435 435	21	26	2.66	.65	1.19	1.20	.52	.66
436 436	11	22	-.58	.57	.84	.72	-.23	-.29
437 437	17	20	3.49	.83	.98	2.64	.19	1.40
438 438	10	14	-.84	.70	1.10	1.17	.36	.47
439 439	10	25	-2.74	.52	.58	.58	-1.15	-.72
440 440	18	22	3.11	.72	.67	1.59	-.48	.99
441 441	5	14	-6.27	.80	.80	.74	-.35	-.22
442 442	10	19	-1.02	.56	2.78	3.09	3.18	2.73
443 443	9	19	-.21	.57	.87	.62	-.20	-.51
444 444	6	10	-4.04	.87	1.30	1.20	.69	.50
445 445	7	19	-2.36	.60	.98	.99	.11	.18
446 446	12	21	-1.22	.52	1.17	1.13	.56	.41
447 447	23	26	3.77	.78	.26	.14	-1.58	.42
448 448	11	29	-2.78	.52	1.33	1.28	.95	.65
449 449	15	29	-.71	.46	.86	.83	-.37	-.24
450 450	12	25	-.19	.49	.87	.68	-.28	-.49
451 451	18	23	2.07	.63	.81	.60	-.25	.08
452 452	15	20	2.51	.68	.55	.39	-.88	.20
453 453	22	37	.28	.47	2.17	1.81	2.50	1.43
454 454	21	35	2.25	.56	1.82	1.43	1.68	.72
455 455	16	20	3.01	.73	.52	.47	-.81	.52
456 456	5	14	-3.54	.77	.73	.63	-.35	-.33
457 457	6	10	-4.04	.87	1.07	.97	.31	.22
458 458	21	26	4.41	.59	1.06	.51	.28	1.09
459 459	5	10	-4.80	.88	1.01	1.03	.22	.30
460 460	22	24	4.35	.92	.34	.13	-1.11	.99
Mean			-1.32		.96	1.09	-.08	.26
SD			3.44		.53	1.79	1.08	1.13

The columns in Table 31 are similar to those for the item statistics in Table 29 above, except that there are no different thresholds. SCORE shows that actual test score and MAXSCR the maximum possible score for each candidate, as calculated by QUEST.

The ESTIMATE is the ability estimate, in logits, for each candidate, with a standard error estimate. At the default probability level of .50 used in these analyses, a candidate with an ability estimate of (say) 0 would have exactly 50% chance of succeeding on an item with a difficulty level of the same value. Stronger candidates will have positive values for ability level, weaker candidates negative values.

The last four columns show FIT statistics, the INFIT MEAN SQUARE, the OUTFIT MEAN SQUARE and the standardised INFIT t and OUTFIT t values. The better a candidate fits the model, the closer the mean squares are to a value of one and the t-values to a value of zero.

Case misfit

Using the formula quoted in the previous section for the range of fit as the mean plus or minus twice the standard deviation of the infit mean square statistic (McNamara 1996), acceptable values for the cases in this analysis would be from -0.10 to + 2.02 (based on an infit mean square mean of .96 and a standard deviation of .53, reported in Table 28 above and also the bottom row of Table 31).

A total of 21 cases (candidates) reported in Table 31 have infit mean squares above +2.02, and none have values below -.10. This represents 4.6% of the total of 460 candidates whose tests scores were included in the IRT analysis. This is a worryingly high figure, given McNamara's suggestion of 2% as an acceptable limit: "A test which produces significant levels of person misfit (greater than 2 per cent of candidates) suggests that it is unworkable as a measurement procedure for too many candidates, and

will need revision to reduce this number" (McNamara, 1996: 178). Given that the data for this analysis come from the pilot version of the Five Star test, a reasonable target might be set for reducing this proportion of misfitting cases to below the limit of 2% for subsequent versions.

A detailed analysis of this person misfit is beyond the scope of this research, but scrutiny of each individual misfitting case would reveal a series of unexpected responses that might form a pattern over several tasks. Individual learner maps show each candidate's responses to each task attempted, and the learner map for one of these misfitting cases is considered in the following section. Reference back to candidate records might reveal consistent patterns in their educational or language background.

6.3.4 IRT outputs 4: distribution of both items and cases

The item difficulty and person ability estimates can be plotted on a single graph. Such an item-ability map with the distribution of both items and cases on the same logit scale is shown in Figure 5.

The figures on the far left show the logit scale on which both items and cases are calculated and displayed. The xxx s on the left of the centre margin represent the cases (candidate tests) according to their ability estimate; in this analysis, 460 in total. The figures on the right hand side of the centre margin show items plotted according to difficulty estimates, with the postscript .2 or .3 to indicate which threshold (task exit level) is indicated.

Figure 5 QUEST output of item-ability map

Item Estimates (Thresholds)

16/ 6/99 17:51

all on all (N = 460 L = 73 Probability Level= .50)

9.0	X	120C.3
8.0	X	113C.3 110B.3
7.0	X	109P.3 114S.3
6.0	X	107P.2 110B.2
	X	102U.2 109P.2
	XX	76We.3 104K.3 105K.3
5.0	XX	98Fr.3 101U.3 120C.2 101U.2 103C.3
	X	94He.3 98Fr.2 105K.2
	XXXX	97Ti.3
	XXXXXXXXXX	
4.0	XXX	104K.2
	XXX	74Li.3 75Sa.3
	XXXX	56Na.3
	XXXXX	114S.2
	XXXX	60Si.3 68Ne.3 91Ch.2
3.0	XXX	88Ri.3 89Cl.3
	XXXXXXX	55Ku.3
	XXXX	16Pa.3 97Ti.2
	XXXXXXXXXXXX	57Ma.3 65Re.3 76We.2 89Cl.2
2.0	XXXXXXXXXX	29Fr.3 53Tr.3 72Le.3
	XXXXXXXXXXXX	36Si.3 71In.3 94He.2
	XXXXXXXXXXXX	22Sh.3 69Ne.2
	XXXXXXXXXXXX	50Ro.3 74Li.2 92Tr.3 123T.3
1.0	XX	54Po.3
	XXXXXXXXXXXXXXXXXXXX	88Ri.2
	XXXXXXXXXXXXXXXXXXXX	47Si.3 58Sp.3 72Le.2 75Sa.2
	XXXXXXXXXXXXXXXXXXXX	56Na.2 59Pu.2 61Sp.3
	XXXXXXX	36Si.2 62Sp.3
.0	XXXXXXXXXX	55Ku.2
	XXXXXXXXXXXXXXXXXXXX	12Fa.3 15St.3 24Fo.3 47Si.2 123T.2
	XXXXXXXXXXXX	28Si.3 53Tr.2 60Si.2
	XXXXXXX	23Ve.3 25La.3 29Fr.2 54Po.2 63Ro.2
-1.0	XXXXXXXXXXXXXXXXXXXX	14Ad.3 19Re.3 26Ke.3 33Tr.3 65Re.2
	XXXXXXXXXXXX	28Si.2
	XXXXXXX	16Pa.2 51Ro.2 92Tr.2
	XXXXXXXXXXXX	13Al.3 61Sp.2
	XXXXXXX	4Nam.3 58Sp.2
-2.0	XXXXXXXXXXXXXXXXXXXX	7Sch.3 15St.2 22Sh.2
	XXXXXXX	

	XXXXXX	26Ke.2 34Tr.2
	XXXXXXXXXX	17Re.3 33Tr.2
-3.0	XXXXXXXXXXXXXX	10Sc.3 19Re.2 25La.2 62Sp.2
	XXXXXX	50Ro.2
	XXXXXXXXXX	11In.3 24Fo.2
	XXXXXX	23Ve.2
	XXXXXXXXXX	
-4.0	XXXXXXXXXXXXXX	14Ad.2 30Re.3 57Ma.2
	XXXXXX	
	XXXXXX	31Wr.3
	XXXXXXXXXX	13Al.2
-5.0	XXXXXXXXXXXXXX	
	X	17Re.2
	XXXXXXXXXXXXXX	12Fa.2
	XX	
-6.0	XX	
	XX	10Sc.2
	XXXXXXXXXXXXXX	
	XXXX	4Nam.2
-7.0	XXXXXX	5Bas.3 11In.2 30Re.2
	XXXX	7Sch.2
	X	
	XXXX	
-8.0	XXXXXXXXXX	
		6Sch.2 8Bas.2
-9.0	XXXX	
-10.0	XXX	27Wr.2
		5Bas.2
-11.0		
-12.0		31Wr.2

Each X represents 1 students
Some thresholds could not be fitted to the display

— $\frac{1}{2}$ —

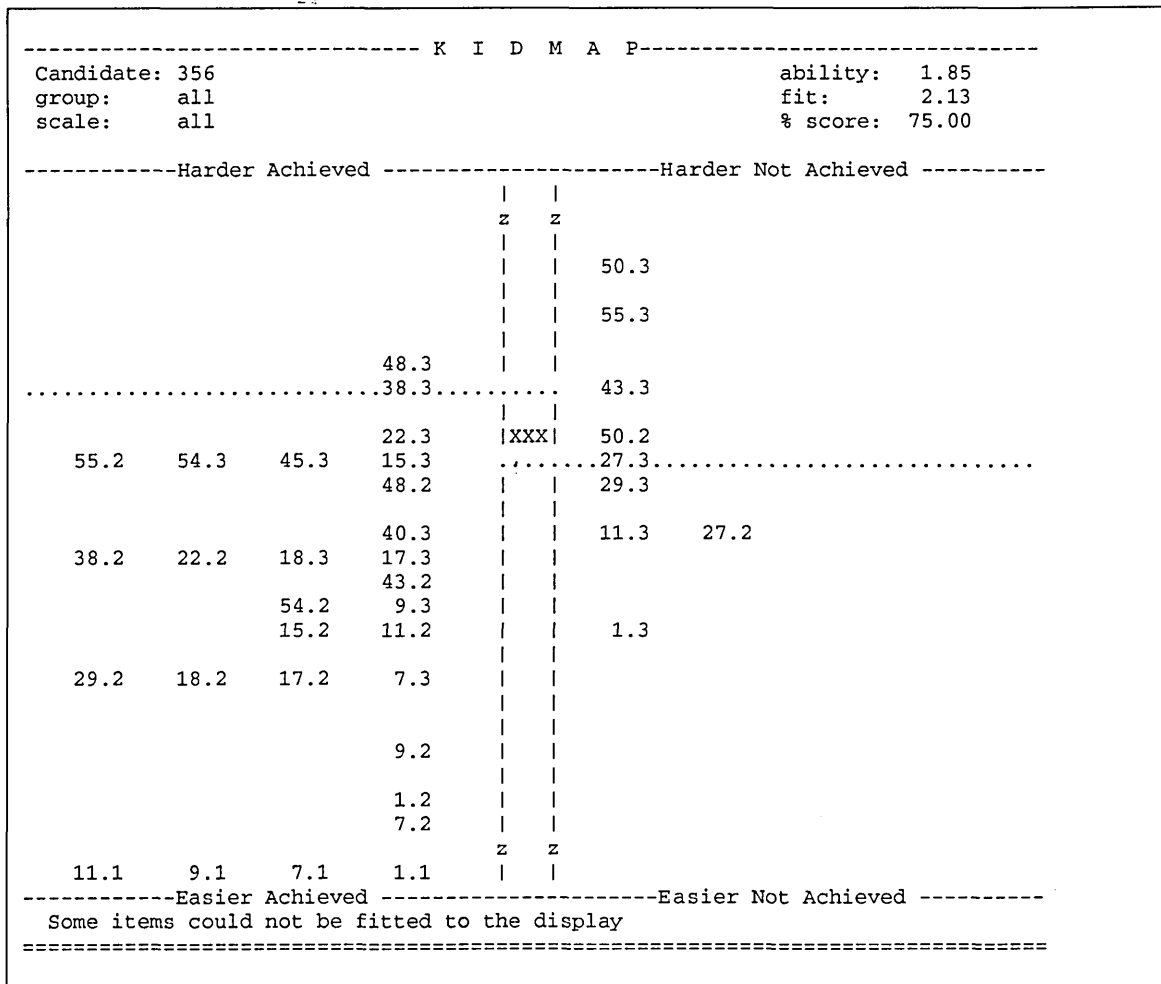
the top left or bottom right quadrants, i.e. which were supposedly too difficult but he nonetheless got right, and those which should have been too easy yet he got them wrong. In this case, only task 40 falls outside the error interval, in the bottom right hand quadrant; according to the model, he could have been expected to perform this task satisfactorily at the middle exit level (40.2) but not at the higher exit level (40.3). In fact, he did not reach even the middle exit level. The top exit level for task 19 (19.3) is just on the upper limit of what the model predicted he might be able to complete, and he did so.

We can compare this learner map with one for another candidate who the case estimates indicated was misfitting (Figure 7).

³ for the purpose of reporting IRT results, only the first serial number of each task is used here, so that task 1-4 *Names* is reported as task 1. The second digit .1, .2 or .3 refers to the three exit levels from each task, corresponding roughly to low, medium or high performance.

Figure 7

Individual IRT learner map for candidate 356



The overall ability level estimate for candidate 356 is 1.85 (shown in the top right-hand corner of Figure 7), well above the mean for the whole sample of -1.32 (from Table 28). However, there are five tasks on which the model was unable to predict his performance correctly. Four of them are in the bottom right-hand quadrant: for tasks 1, 11 and 29, he should have performed at the top exit level (1.3, 11.3 and 29.3) but didn't; and according to the model he should also have completed task 27 at the middle exit level (27.2) but failed to do so. Task 48, on the other hand, should have been just above the upper limit of his ability level, but he completed it at the top exit level 3, so it appears in the upper left-hand quadrant.

6.4 Comparison of results between data sets

The availability of data on each task from two quite different sources provides the possibility of triangulation and allows cross-checking of implications from one source to the other. Two examples of this triangulation for Five Star data can be seen in the estimates of task difficulty and item fit.

6.4.1 Task difficulty

The two data sets allow comparison of estimates of item difficulty from two distinct sources, the expert panel's estimate of each task on the ESU 9-point scale and the IRT output of item difficulty on the logit scale. For the latter measure, two separate values are available for each item, as reported in Table 29, because there are separate estimates for each threshold (exit 2 and exit 3 from each task). The panel instructions however did not ask panelists to distinguish between the exits, so only a single score was produced for each item, as reported in Table 19.

Table 32 compares the difficulty ratings for each task from the panel exercise in Table 19 with the IRT analysis in Table 29.

Table 32 **Comparison of item difficulty estimates between data sets**

Item / task	Panel rating	QUEST Exit 2	QUEST exit 3
4 Names	1.3	-6.50	-1.77
5 Basnum	1.8	-10.26	-6.91
6 Schl1	2.3	-8.28	.
7 Schl2	2.9	-7.31	-2.02
8 Read	2.1	-8.28	.
10 Schl3	3.5	-6.13	-2.87
11 Intnm	3.0	-6.91	-3.24
12 Famil	3.7	-5.53	-.40
13 AlHar	3.7	-4.56	-1.67
14 Advnm	4.6	-4.06	-1.00
15 Strep	4.1	-1.94	-.25
16 Paper	4.1	-1.50	2.29
17 Read3	2.8	-5.38	-2.71
19 Read4	3.8	-3.00	-.89
22 Shape	4.1	-2.03	1.50
23 Vehic	4.1	-3.47	-.78
24 Footb	4.3	-3.25	-.39
25 Ladde	4.3	-2.84	-.68
26 Kettl	4.3	-2.47	-1.02
27 Writi	.	-9.86	.
28 Signs	4.5	-1.13	-.59
29 Fridg	4.9	-.75	1.81
30 Read2	2.5	-6.81	-3.92
31 Writi	.	-11.57	-4.38
33 Traf1	3.3	-2.69	-1.00
34 Traf2	3.6	-2.42	.
36 Sign2	4.4	.06	1.67
47 Sign3	4.5	-.28	.66
50 Road	4.0	-3.06	1.15
51 Road2	4.6	-1.46	.
53 Train	5.3	-.50	1.92
54 Popul	5.5	-.84	1.09
55 Kuwai	6.0	-.03	2.45
56 Nagor	6.4	.28	3.30
57 Tea	4.2	-3.94	2.17
58 Spec1	5.4	-1.72	.64
59 Punct	5.0	.44	.
60 Singa	5.2	-.44	2.99
61 Spec2	5.0	-1.63	.41
62 Spec3	5.0	-2.97	.20
63 Accid	5.3	-.66	.
65 Regio	5.8	-.94	2.13
68 News1	5.3	-.81	3.04
69 News2	5.4	1.36	.
71 Instr	5.6	-.66	1.74
72 Leban	6.6	.47	1.79
73 Inst2	5.9	.	.
74 Lille	6.0	1.19	3.69
75 Saudi	5.5	.50	3.55
76 Weath	6.3	2.16	5.26
88 Riyad	5.3	.78	2.70

Table 32 (continued) Comparison of item difficulty estimates between data sets

Item / task	Panel rating	QUEST Exit 2	QUEST exit 3
89 Clima	5.7	2.13	2.68
91 Child	6.1	2.89	.
92 Trave	4.9	-1.47	1.31
94 Heath	5.4	1.53	4.36
97 Tim S	6.7	2.38	4.31
98 Free	5.8	4.47	4.97
100 Unit	7.0	.	.
101 US H	7.0	4.69	4.97
102 UNID	6.6	5.49	.
103 Prio	6.8	.	4.68
104 Karo	6.1	3.88	5.42
105 Kar2	5.9	4.44	5.27
107 Pric	6.8	5.95	.
108 Prod	6.5	.	.
109 Pors	6.7	5.59	6.72
110 Book	7.0	5.84	7.60
112 Bosn	6.9	.	.
113 Cons	7.5	.	7.88
114 SA r	6.7	3.09	6.51
115 Hone	6.4	.	.
120 Cons	6.9	4.91	8.81
123 Comp	5.2	-.22	1.26

The Panel ratings are the mean of the 12 panellists' judgements of each task difficulty from Table 19 made on the 9 band ESU scale in Table 11. The QUEST scores are the difficulty scores from Table 29 calculated by the IRT programme QUEST for the second and third exits (medium and high performance) for each task.

Correlations between these difficulty ratings in Table 32 were calculated using SPSS, and the results are shown in Table 33.

Table 33 Correlations between IRT and panel estimates of task difficulty**a) Parametric: Pearson product-moment correlations**

	Panel ratings	Quest exit 2	Quest exit 3
Panel ratings	1.000	.918 **	.890 **
N	71	64	56
Quest exit 2	.918 **	1.000	.939 **
	64	66	55
Quest exit 3	.890 **	.939 **	1.000
	56	55	57

** correlation significant at the 0.01 level (1 tailed)

b) Non-parametric: Spearman's rho (rank correlation)

	Panel ratings	Quest exit 2	Quest exit 3
Panel ratings	1.000	.918 **	.887 **
N	71	64	56
Quest exit 2	.918 **	1.000	.926 **
	64	66	55
Quest exit 3	.887 **	.926 **	1.000
	56	55	57

** correlation significant at the 0.01 level (1 tailed)

Being an interval based statistic, the Pearson product-moment correlation assumes an equal distance between the points on the scales, and it might be prudent not to make this assumption of the panelists' use of the ESU 9-point scale, and in principle to rely instead on the Spearman rank correlation coefficient. In practice, there is very little difference between the two sets of statistics. All the correlations shown are significant at $p = 0.01$.

6.4.2 Task-to-test fit

This exercise takes the tasks whose estimates of item fit fall outside the suggested range of fit (0.35 to +1.47) as discussed in 6.3.2 above and looks for further information from scrutiny of the tasks and the expert panel data set for possible explanations.

Table 34 Panel consensus for misfitting tasks

Task number and name	Threshold 2	Threshold 3	Infit mean square	Infit t	Panel skills allocation	Panel difficulty estimate on 9-band scale
3- 6 School/study 1	-8.28	-	0.34	-.9	Listening, speaking	2.3
5-8 Basic Reading	-8.28	-	0.34	-.9	Reading	2.1
21-28 Signs	-1.13	-.59	1.76	4.2	Reading	4.5
27-36 Signs2	.06	1.67	1.79	1.9	Reading	4.4
28-47 Signs3	-.28	.66	1.41	2.9	Reading	4.5

The first two tasks in Table 34, 3-6 School/study 1 and 5-8 Basic Reading, were considered in 6.3.2 above as misfitting through poor discrimination, based on the very low item difficulty values of -8.28, compounded by a small sample size. This interpretation is reinforced by the panelists' mean difficulty estimates of 2.3 and 2.1 respectively, on the 9-level ESU scale used as a yardstick (Table 11). At level 2, a candidate 'does not really have sufficient language to cope with normal day-to-day real-life situations'. Only two of the 73 working tasks have lower mean panelists' ratings.

It is noticeable immediately that three remaining misfitting tasks 21-28, 27-36 and 28-47 are of the same type, where the candidate sees a series of messages displayed for a timed period and has to select the message that is appropriate for the context given. The contexts are a road sign prohibiting heavy goods vehicles (task 21-28); a warning sign

on an airport perimeter fence prohibiting photography (task 27-36); and a poison warning on the label of a bottle (task 28-47) The rubric for the interviewer that is common to all three items reads:

"Click on the Arabic instructions which tells the candidate that one of the texts (1 to 6) is the written message on the sign in the picture. He will have five seconds to silently skim-read each text and select the best choice. This sequence may be repeated once. To gain the maximum score, the candidate's correct selection should be made at this point. The candidate who make a correct selection only after a repeated sequence and reviewing one or more individual texts will gain an average score".

The candidate is given these instructions in a recorded message in Arabic. The overall criterion for scoring is 'Text identification' and the three exits are labeled [blank], 'Average' and 'Maximum' but there are no individual pop-up descriptors for each of the three exits. Two of the panelists' comments on these three sign tasks were

- "Procedure could cause embarrassment for slower candidates"
- "I like the authenticity of the signs and the semi-objectivity of the marking. The implicit assumption is that 'faster reading = better reading'. Speeded tests create anxiety!"

Three possible explanations for these three misfitting items suggest themselves.

Hypothesis 1: cognitive load under pressure of time

There is a fixed five second display for each of the six possible signs, and although there is a repeat option, this incurs a penalty in the scoring system, which is based on

the identification of the single correct sign for that context. There is therefore a cognitive load requiring memory and comparison of a series of authentic texts of a specific type under external pressure of time. One hypothesis might be that this is tapping into skills or abilities that are quite distinct from the language skills drawn on by other tasks.

Hypothesis 2: comprehension in the mother tongue

Another possible interpretation of these misfitting items is that the tasks require comprehension of some quite complex instructions in Arabic. While there are several other tasks that also have Arabic instructions yet do not apparently misfit the IRT model, these three have an identical Arabic-language instruction, and no other task shares this rubric. It may be the Arabic instruction itself that is at fault, rather than the task. It might be also worth looking at the item-response pattern of other tasks with Arabic instructions to see more generally if there is a factor of comprehension in the candidates' mother tongue creating an additional source of variance in their second language performance.

Hypothesis 3: objective scoring

A third possible interpretation is that these tasks are among those with the most objective scoring systems, while the great majority of tasks require more subjective impressionistic judgements to be made the interviewer, using labels such as partial, complete, acceptable, clear, moderate, excellent. A hypothesis here could be that objectively-scored tasks while not necessarily testing different skills are reflecting candidate's language abilities in a way that is not consistent from the other tasks, thus

showing up as misfitting against the IRT model which is based on the dominant task type, subjective scoring.

In each case, further research would be needed to explore the hypothesis, and in particular to seek to explain why other tasks which apparently share the same features, such as a requirement for Arabic comprehension in the case of hypothesis 2 or objective scoring in the case of hypothesis 3, do not also show up as misfitting.

6.5 Summary

Analysis of the results of the two data collection exercises give two different perspectives on the test. The entire panel exercise consists of a very large number of individual complex judgements, and the panelists by and large show themselves able to reach substantial agreement on the skills and levels of proficiency required for each task, and on the type and impact of interaction strategies used by candidates. It is worth emphasizing that at no time did the panelists meet or discuss their judgements or comments face-to-face; these were submitted and circulated entirely in writing.

The stage 1 skill allocation results reported in Tables 12 and 13 accord with one's intuition that in everyday life speaking and listening are likely to be combined at least with each other, whereas reading is much more likely to occur alone; and in so doing contributes to the construct validity of the test. Study skills remains more difficult for panelists to agree on, perhaps because of its uncertain status as a language skill.

Interaction is clearly seen to be important, but again this vindication of its contribution confirms our intuition and helps to validate the construct of the test, but does not tell us how to treat interaction vis-à-vis other skills in research design.

With the panel data all being reported for two sub-groups as well as for the whole group of 12, there is a lot of scope for further investigation of the reliability of panel judgements in these circumstances. Broadly speaking, the results between sub-groups compare well.

The results also allow scrutiny of individual panel members' scoring patterns, and as might be expected, individual panelists show some patterns of consistent divergence from their colleagues, and this can easily be identified and quantified, for example, in the variation in panelists' ratings of candidates' proficiency. In a panel exercise with an extended lifetime and repeated cycles of activity, these could be reduced either through a deliberate focus on moderation exercises or through simply recycling the results at each stage and inviting panelists to review their judgements and resubmit in the light of the panel's provisional consensus.

It is reasonable to expect that the data also reflects an element of random variation in judgements, but this would be harder to find evidence of because it is non-systematic.

The results of the IRT treatment by contrast are based purely on test results, and there is no element of subjective judgement in the analysis. The results themselves are however the outcome of subjective judgements made by the test interviewer, and again a note of

caution should be sounded that all the tests in this analysis were administered by a single interviewer.

The QUEST programme provides a range of outputs for analysing task and candidate performance, and the examples given here do not exhaust its capabilities. The fit statistics in particular provide valuable information in a number of ways. They help to validate individual tasks and to highlight those that are significantly misfitting; scrutiny of the tasks, allied with panelists' ratings and comments, go some way to elucidating the possible causes of the misfit. That three of the misfitting items are of one particular type is a strong indication that whatever it is they are measuring is not the same underlying trait as the rest of the test.

The fact that only five of 73 tasks misfit overall is evidence for overall validity of the test. The fact that nearly five percent of the candidates are misfitting is greater cause for concern, and careful monitoring of this would be needed in an extended programme of data collection and analysis to identify whether in fact the IRT model was for some reason inappropriate for as many as one in 20 of the candidates in this population.

Simply re-routing the algorithm to use some of the tasks more than they have been in this dataset would provide much more information for the IRT analysis to work on, and a series of minor changes to the test could be implemented and analysed for impact. Such incremental changes are technically quite feasible given the ease of programming and 'open architecture' of the test, but raise more fundamental questions about the model for validation of a continually changing test that are addressed in the next two chapters.

In chapter nine, we come back again to apply the model for continuous validation to the

Five Star as an exemplar of the broad type of test to which this might apply.

- 7.0 Introduction
- 7.1 Summary of current validity issues
 - 7.1.1 Componential vs. unitary views of validity
 - 7.1.2 Empiricist vs rationalist validity
 - 7.1.3 Internal vs external validity
 - 7.1.4 Evidential vs consequential validity
 - 7.1.5 Product and process in validation
- 7.2 The current pedagogical approach to teaching and testing:
 implications of the communicative approach
 - 7.2.1 Direct oral tests
 - 7.2.2 Interaction
 - 7.2.3 Task-based structure
 - 7.2.4 Individualisation
 - 7.2.5 Topicality
 - 7.2.6 Authenticity
- 7.3 Summary

7.0 Introduction

This chapter first considers some of the theoretical distinctions in the discussion on validity and test validation, and then applies them critically to the key features of communicative tests that have emerged in previous chapters. It attempts to tease out the practical implications of these validity and methodology issues and leads into the

formulation of a theoretical model for continuous validation in the next chapter.

Chapter nine then applies that model specifically to the Five Star test.

7.1 Summary of current validity issues

This section summarises and comments on the major validity issues raised in previous chapters, and examines their implications for context-sensitive, adaptive tests of communicative performance. These validity issues are overlapping perspectives rather than mutually exclusive models.

7.1.1 Componential vs. unitary views of validity

Validity was viewed historically in the literature (see chapter three) as being componential, and distinct types of validity were to be evidenced separately. The current paradigm is to see validity as a single construct; while diverse types of evidence can and should be sought, it is accumulated to establish validity as a unitary concept. There is no preferred range or combination of sources of evidence; quality of evidence is primary.

At the same time, the range of possible sources is greater, particularly with the addition of consequential considerations, and the number and recurrence of data collection activities is greater, with a focus on recurrent process rather than one-time event. This creates a greater onus on the test developer to plan the data collection and synthesize the

validity findings than previously, where distinct sub-types of validity might be reported separately. The choice of sources of validity evidence will be heavily context-dependent, with commercial and ethical considerations also likely to influence decision-making.

The focus of this research is primarily on test validation, but applied to real-world programmes, of which the language testing may only be one component, rather than purely academic test development. There is therefore a requirement for a wider dimension including practical as well as theoretical considerations to evaluate the overall success or failure of a test, and the term 'programme evaluation' seems appropriate for this. As quoted in section 3.2.6, 'Usefulness' has been proposed as an overarching criterion that combines different test qualities: "Usefulness = Reliability + Construct validity + Authenticity + Interactiveness + Impact + Practicality" (Bachman and Palmer, 1996: 18) and programme evaluation should include the broader ambit of usefulness as well as the narrower focus on test validity.

7.1.2 Empiricist vs rationalist validity

As ever, there is a potential for tension between approaches which are quantitative, statistically-based and narrowly-focused empirical studies and those which are qualitative and more broadly-based naturalistic data collection exercises. The current consensus would see both types of data as being necessary to provide a balanced synthesis of the different types of evidence. There is a danger of this contrast being seen as offering only two diametrically opposing options; in reality, qualitative data can and

should be elicited in a systematic manner, and the value of quantitative data is dependent ultimately on the quality of the subjective interpretation of it.

In language programme evaluation more widely, as well as in the validation of language proficiency tests specifically, naturalistic methods have struggled in the past to be considered scientific and of comparable value to 'hard' empirical data in the positivistic paradigm. A particular problem has been to find ways to elicit stakeholders' views - ultimately, personal opinions - in a sufficiently objective manner, so that the variance caused by the subjectivity of the survey methods does not drown out or distort the underlying data. Group decision-making techniques such as Delphi offer a way to collate a consensus of group opinion without compromising the richness of the data collected.

A carefully constructed panel exercise can be used to generate qualitative data that can be focused on the key questions most relevant to the test development at a particular time, and these qualitative data can be analysed using quantitative means to maximise the consensus view while minimising the peer-group's influence as a source of subjective bias. Electronic media may be particularly appropriate for this. The expert view, however it is elicited, generates rich data on internal and external content validity, comment on face validity, and comparison of individual tasks against the core construct.

Unlike empirical data based on the analysis of test scores, some types of panel activity can be 'case free' and independent of actual test events, and can even begin before the accumulation of the large quantities of test data that are needed for IRT.

Traditional test statistics suffer from being sample dependent, and therefore of questionable transferability to other samples or a wider population. Most such statistics, such as item discrimination, item facility, factor analysis, and common internal reliability measures such as Cronbach's alpha and Kuder-Richardson formulae 20 and 21 (Crocker and Algina, 1986) are in any case inapplicable to the incomplete data sets generated by adaptive tests. Comparison by correlation against criterion tests is widely viewed in the literature as the archetypal form of empirical validity and is feasible for adaptive tests, but the difficulty of interpreting a typical external correlation is discussed below.

IRT statistics appear to offer a solution to both these constraints, by generating sample-free item estimates on a stochastic basis, and by operating comfortably on adaptive test data. They do, however, require larger data sets, which exacerbates the problem that empirical data can only be collected once a full working pilot form of the test has been developed, by which time major assumptions about constructs will already have been made.

As well as providing independent sources of validity evidence, expert panel and IRT data support and inform each other and provide invaluable triangulation evidence. An obvious example is item facility, with estimates of the language proficiency level required for each task by an expert panel and IRT sources of data being compared and correlated.

Another example of triangulation between expert panel and IRT data relates to content and construct evidence. The panel judgements on each task can be focused on its

communicative qualities, for example, while the fit statistics generated by IRT measure how that task measures up against the criterion being measured by the test as a whole, and they thus provide two perspectives on the construct validity of each task. The empirical fit statistics can be interpreted as a direct measure of the extent of construct underrepresentation or construct irrelevance (APA, 1999), but not be able to discriminate between them. On the other hand panel data, focused for example on the percentage breakdown of skills tested by each task, would indicate whether the misfit was due in the experts' opinion to a task that was too narrow or too broad, whether there were obvious reasons for the misfit and by implication easy ways to improve the task, and also the extent to which the panel themselves were able to reach agreement on these questions.

In many cases, it is not clear what conventional sub-categories of validity or reliability the data collected falls into. In the first example of panel/IRT data complementarity above, the extent of overlap between panelists' and IRT estimates of task facility, a correlation coefficient between these two could be seen as an indication of internal reliability, or of validity of test data against the external criterion of panelists' ratings. The second example could be interpreted as content or construct (or to an extent face) validity.

Traditionally, this ambiguity might be seen as a failure of operationalisation of the validity construct in a paradigm with an idealised, not to say simplistic, view of the nature of language proficiency and a strictly componential view of validation. In reality, the confusion has always been there, particularly in respect of the interpretation of statistics such as correlations. A unitary approach to validation allows us to take a more

enlightened perspective; we seek validity evidence from multiple sources, and do not need to categorise each type in such a rigid framework in order to value its contribution to the overall task of validation.

7.1.3 Internal vs. external validity

This distinction cuts across the qualitative / quantitative contrast, and allows both internal and external validity to be of either type. Internal validity sources would include item statistics as well as face and content validity considerations. External validity allows expert comparison against external criteria as well as empirical data such as correlation studies.

The central dilemma of external validity is that each test is unique, in its combination of content, aims, target group and context, or else there would be nothing to distinguish it and no motivation to develop it. To the extent that a test diverges in any important respect from existing criterion tests, any interpretation of external comparisons will be contentious. Evidence from similar tests in other settings will always be valuable but will need to be interpreted with caution.

Added to this, the greater concern for the often divergent objectives of the different stakeholders typically found in a local context entails the likelihood of ambiguity in the interpretation of such comparisons. A culturally sensitive perspective must respect the possibility that different stakeholders will retain incompatible and deeply-held values which necessarily lead to differing interpretations of external validity data.

7.1.4 Evidential vs. consequential validity

This contrast also cuts across the qualitative/quantitative distinction, and anticipates the product/process distinction below. Data of different kinds for evidential validity is collected about the test and the test results, and may be internal or external, but is restricted to the test and its immediate context, including criterion-related comparison. Consequential validity is largely based on external data, and it opens up a much wider perspective, to ask how the test scores are used.

First introduced by Messick (1988), referred to in 3.2.6 above, this emphasis on validity as attaching to the interpretation of test scores rather than tests in themselves is now pivotal to the current validity paradigm (e.g. APA, 1999). It raises important ethical questions over the control and authority of test use. Evidential validity is relatively straightforward, in terms of data collection, although the interpretation may be open to discussion. Consequential validity requiring the use and interpretation of test results to follow strict guidelines takes at least some of responsibility out of the hands of the test developer, and where the use a test is put to differs significantly from that originally supported by the developer, the test user bears the burden of continuing the task of test validation in the new context.

The term 'impact' is used by Bachman and Palmer to embrace the various possible consequences at a 'macro' level of society and the education system as well as a 'micro'

level of individuals (Bachman and Palmer, 1996: 29). Impact also includes the washback effect of testing on an associated teaching programme.

Where difficulties in collecting consequential data can be anticipated, the developer can seek an undertaking to restrict test use only to certain target groups and contexts for which validity evidence has already been collected, but in an extreme case this may signal a good intention rather than a contractually enforceable reality. As is commonly the case for computer software a system of licensing, rather than outright sale, of the use of the test may provide some partial recourse and retention of control. This situation might arise where there is commercial pressure to maximise the sale or use of a test, diverging cultural interpretations of the extent to which adherence to such an undertaking imposes, or political decision-making processes that are outside the test developer's control. Ironically, in a strongly hierarchical situation of use such as a governmental, military or commercial/industrial context, the wider the initial support for the development or adoption of a test the greater the likelihood of subsequent policy decisions overriding technical concerns about test validity and the interpretation of scores.

Such a concern may arise where a test has humble origins in a modest, small scale development for a very specific local context, such as a single training programme, with severe limitations on the human technical resources available for comprehensive validation, especially empirical. Adoption on a larger scale, even with the same target population, may not be justified by the existing validity evidence, yet the stakeholders may not appreciate why this is so. Another example of over-extension occurs where the extent of test validation can reasonably justify routine everyday decisions - class

placement or diagnosis, for example, which can be confirmed or rectified subsequently - but cannot justify the same test being used for crucial career or academic admissions decisions (so-called 'high stakes' testing).

Even where complete control over the use of a test remains with the original developer, new candidate groups may emerge that differ in terms of such variables as age, nationality, purpose or background, and at some point the differences will be such that the test needs to be effectively revalidated for this new target population. Only the continuing collection, analysis and longitudinal monitoring of data can allow the developer/user to determine when that point is reached. Particularly for communicative tests where topicality is a fundamental requirement of the construct, what is current affairs for one year-group of teenagers or cohort of young adults is history for the next.

The diversity of possible test users and their interests may mean that different levels of support are required by the test developer to enable them to interpret test results properly. For a given test, the test user might be the individual candidate and/or his/her family sponsor; the admissions tutor of an academic institution; the training or recruitment manager of a potential employer; or the test administration staff of a licensee or purchaser of the test. In each case, consequential validity requires that, while the developer is responsible for providing evidence and a rationale for the intended test use, it is the user who is ultimately responsible for evaluating the evidence in a particular context. Only the user can know the implications of any decision that draws on the test result, such as in a recruitment or admissions procedure.

This requirement on the test developer to support the interpretation of scores by end-users creates an immediate dilemma in deciding how much information should be provided, balancing clarity and brevity against precision. On the one hand, the greater the level of detail, and the more real-life contingencies that can be anticipated, the fuller the level of support that is provided. On the other hand, non-specialist test users (which is most of them) will be put off by more information than they need, and will simply look for a table of equivalence or set of performance descriptors that they can interpret with ease.

Even professional training managers and academics seek simple answers, since they have to make 'yes' / 'no' decisions under pressure of time, and will prefer a single number score or scale to a profile containing multiple descriptors.

7.1.5 Product and process in validation

This contrast follows directly on from the evidential vs. consequential bases for validity above. What was formerly the end of the task, the one-time collection of sufficient evidence to establish the independent validity of a test once and for all, is now only the first stage. There is inevitably a tension between validation as a necessary pilot stage and a continuing process; in its everyday general meaning, a document such as a passport or licence cannot be issued or used unless it is 'valid' and this 'validation' is a one-time activity. What is required here is an important change of perspective, involving tests users being made aware of and accepting their responsibilities for the interpretation of results.

One of the implications of consequential validity as a continuing process is that you can never reach a point where the test can be declared 'fully valid', and therefore that at any point along the way, the evidence for validity can only ever be partial and incomplete. While validation must be a continuing process, there must in practice be waystages for summative evaluation of the process to date and there also needs to be a procedure by which the validation itself can be evaluated on a continuing basis. It will often be the case that influential stakeholders will continue to seek simple answers to simple questions such as 'Is the test valid?' and may have little patience for answers that are heavily-qualified with conditions. They may also resist the idea that further large-scale and therefore costly data collection is necessary for validation when minor changes are introduced to a test or when it is applied to a slightly different candidate population. The model for continuous and cyclical data collection proposed in the next chapter may in fact provide a form of insurance; if accepted by stakeholders and routinely implemented from the outset, it may obviate the need for special data collection activities.

Both the evidential vs. consequential and product vs. process distinctions may generate data that fall strictly outside the scope of validation but are relevant to the broader activity of programme evaluation.

7.2 The current pedagogical approach to teaching and testing: implications of the communicative approach

The communicative approach is currently the dominant paradigm in English language teaching and testing. To accept the fundamental premises of the communicative approach outlined in chapter two is to accept that testing of performance is essential, where the emphasis is on demonstrating the mastery of practical skills rather than the traditional pen-and-paper type of test, and a number of specific implications for test content and structure follow on which need to be examined in the light of the validation issues raised above.

7.2.1 Implications of the communicative approach 1: direct oral tests

The distinguishing characteristic of a performance test is that it involves some kind of judgement being made of performance against some kind of scale. For a language performance test, the behaviour being rated must be either in speech or writing; from the 'work sample' approach (McNamara, 1996), it is that behaviour that is itself the object of assessment, from the cognitive approach that elicited language sample is the means to get at the underlying cognitive competence.

For both speech and writing, the two major test facets that concern the test developer are therefore the raters who make the judgements, and the scale they use to do so. To validate the raters, test developers/ users traditionally use a variety of procedures generally described as moderation. Typically, this involves group events where

individual raters make judgements of sample tests, pool their ratings, compare individual ratings with the consensus, and then negotiate as a group to identify and agree the characteristics of the performance that define the particular rating score. Such a procedure for oral test moderation is described in Underhill (1987).

Validation of scales is carried out using similar data, analysed from the perspective of consistency of judgements made by raters at each level of each scale. Identification of the points on each scale where raters have greater difficulty achieving consensus is followed by a collective analysis of which features of the scale are ambiguous or difficult to interpret.

The major problems with this approach to rater and scale moderation are

- a) the moderation event is an artificial test situation, and assumes that rater behaviour at such events accurately reflects behaviour in 'real-life' tests
- b) the collection of consensus is often carried out in a non-systematic and non-anonymous manner, making it likely that there will be a significant influence on behaviour by the peer group or dominant individuals
- c) for oral tests, the test samples used are likely to be recorded, which immediately puts them at one remove from the direct, live oral test event that is being validated

Alternatively, rater and scale moderation can be based on the analysis of results of each individual rater's judgements over many test administrations, and comparison of such statistics as the mean, median and distribution of scores awarded. Since these are based on different candidates and often on candidate groups from different test centres or contexts, it may be difficult to compare rater behaviour with confidence.

An additional feature of normal spoken language is that the communication between participants takes place in real time, with both or all participants needing to be present together. Communication via writing, on the other hand, can be deferred yet still authentic, with a time gap of any duration between the encoding by the author and the decoding by the reader. While performance tests of writing share with oral tests the concern for validation of both raters and scales, and some of the other implications of the communicative approach listed below, it is much easier to duplicate in a test context the deferred communication model characteristic of real life, by asking the candidate to complete a task which is evaluated subsequently.

Some of the issues that this real-time live testing introduces are construct-related, such as the nature of interaction, and this is discussed below. Other issues concern the practicalities of test administration, but they have major implications for test design and economics.

a) Most oral elicitation activities require interaction between candidate and an interlocutor. In the terminology used here, an interlocutor engages a candidate in oral interaction, but does not assess; an assessor rates performance against a scale, but does not interact; and an interviewer combines both roles. The combination of both roles imposes considerable strain on an interviewer; it is in fact very difficult both to engage a candidate in remotely realistic conversation and at the same time to evaluate his or her performance. On the other hand, separating the roles of interlocutor and assessor doubles the human resource requirement and cost for each event, as well as imposing a larger audience on the candidate (and there may in additional roles concerned with test administration).

- b) Even with a single interviewer, there is a very much greater cost incurred in getting candidates and interviewer together in the same place at the same time.
- c) Furthermore, many test systems (for example, the UCLES main suite and Trinity College Grade Exams) take longer for higher levels of proficiency, with consequently greater administration costs for tests for more proficient candidates; see Table 6 in chapter four.
- d) Spoken language is ephemeral. Where a written sample is easily stored and transmitted or copied for moderation or remarking after the elicitation event has taken place, a spoken sample can only be used in this way if a conscious decision has been taken in advance to record the oral test on audio or video media, which imposes additional resource requirements and an extra strain on the artificiality of the event.
- e) The power imbalance implicit in a one-to-one oral interview is so great that it will inevitably lead to the production of unnatural or 'test-type' discourse. There is some evidence that group discussions reduce anxiety and give individuals more confidence to speak (Fulcher, 1996) but it also raises new worries about the influence of dominant peers in the group, the introduction of new sources of bias and the difficulty of getting comparable samples from each group member.

The alternative to having a live interlocutor is to elicit a real spoken speech sample, in a language laboratory for example, and to record it for subsequent rating; it represents a deferred communication model for speech rather than writing. This is what Clark (1975) termed a 'semi-direct' (also known in the United States as SOPI, see glossary), rather than direct oral test (OPI). Examples of test systems that have adopted this semi-direct format are the Test of Spoken English and PhonePass tests described in chapter four and

the speaking component of ELSA, the English Language Skills Assessment (LCCI, current).

While obviating some of the practical difficulties of direct oral testing, the semi-direct format fails to reflect the central criterion of two-way communication in real time that is characteristic of most real-life speech. Van Lier also considered that "'communicative stress' is much more likely to interfere with performance on a taped (semidirect) than in a face-to-face (direct) encounter" (van Lier, 1989: 493). On a comparative analysis of the interaction in direct and semi-direct tests, Koike (in Young and He, 1998: 69) found significant differences in five of the 12 discourse variables examined. In the direct tests, candidates used a wider range of speech acts, quoted more reported speech and switched into their mother tongue more often, suggesting a more conversational interaction than in the semi-direct tests, where candidates used more fillers and corrected more of their own errors, implying a higher level of conscious attention to their language and, possibly, a higher level of stress.

A comparison of discourse in OPIs and SOPIs in Hebrew found that the SOPIs generated more features associated with written rather than spoken discourse, with a genre 'similar to short essay questions on written examinations and assignments' (Shohamy, 1994: 110). In a critique of the semi-direct speaking component of LCCI's English Language Skills Assessment, Banerjee commented "The test taker is expected to give full sentence answers in all cases even where abbreviated responses are more typical in native speaker speech". A further problem she noted with semi-direct tests that aim to test speaking only is that in the absence of a live interlocutor "candidates who fail to provide evidence of their ability to present an argument might do so because

they have not understood what they are being asked to talk about and do not have recourse to a reformulation of the question" (Banerjee, 2000)

One of the advantages of a direct oral test is the possibility of making more efficient use of the time by including an element of adaptivity, where the interlocutor focuses tasks around the candidate's perceived level rather than posing ones which are clearly too easy or too difficult to add useful information to the assessment. Traditionally the problem here has been to benefit from the greater time efficiency of this adaptivity while controlling for the greater divergence of candidate samples and reliance on individual assessor expertise. There are so many sources of subjectivity and hence possible error variance that some test systems have preferred to pull back from direct assessment to semi-direct or even indirect tests that offer greater reliability.

However, IRT offers a systematic way to moderate and validate raters, tasks and scales, based on 'real-life' data. Again it requires a sufficiently large sample for such multi-facet analysis, but assuming this is available, the results can be transferred from one sample to another.

7.2.2 Implications of the communicative approach 2: interaction

While there are some real-life situations in which speech is used but not in a two-way interaction, such as traditional lecturing or making broadcast announcements, almost all speech events involve interaction between two or more people, and most of the test

tasks listed in Table 6 reflect this. Even test tasks that appear to be one-way communication, such as narration or description are often subsequently broadened out by the interlocutor to generate interaction. This is true to a greater or lesser extent of all direct oral tests. However, dealing with interaction is deeply problematic, both in theory and in practice.

The theoretical problem it poses for test validation is about the nature of interaction, and how it should be treated for the purposes of validation.

Is interaction:

- a) simply to be considered a synonym for spoken communication, and therefore a skill that can be addressed discretely from listening, reading and writing? (the four separately tested skill areas of the Cambridge CCSE examinations were formerly labeled reading, writing, listening and oral interaction - UCLES, 1995)
- b) a combination of two of the traditional four skills: interaction = listening + speaking?
- c) a component of communicative ability that is non-linguistic? (e.g. the Canale and Swain model cited in chapter two)
- d) a combination of some of a candidate's language skills, plus extralinguistic features?
- e) a co-constructed aspect that is contributed to, in different measures, by all the participants? The difference in language sample generated by a candidate in conversation with an effusive interlocutor and with a monosyllabic one cannot be ascribed solely to the candidate

- f) not primarily an aspect of the participants' language proficiency at all, but rather a feature of the event at which it takes place, related perhaps to Firth's 'context of situation'?

In terms of research design, is interaction to be postulated as a variable, and so in principle at least observable and measurable, or is it an underlying construct which can only be approached indirectly through operationalisation to other variables?

If a test aims to operationalise communicative competence, but does not directly address interaction, does this constitute 'construct underrepresentation' (APA, 1999: 10), in that an important part of the construct is not reflected in the variables being measured? Or, on the other hand, if interaction is one of the scales for assessment, is this 'construct irrelevance', on the basis that it attaches to the event and/or to the other participants as well as the candidate?

The influential Canale and Swain model of spoken language ability, illustrated in chapter two, includes as the sub-components of the 'strategic competence' interaction patterns, interaction skills and non-verbal features of interaction, suggesting that they view interaction as a core component of language proficiency but involving more than purely linguistic concerns. The ELI test has as one of its salient features 'interactional facility' and operationalises this to include 'appropriate eye contact/posture' as well as specifically linguistic criteria (ELI 1999, quoted in section 3.3.2 above)

The Bachman (1990) model, also illustrated in chapter two, does not use the word 'interaction' at all. However, in a discussion of authenticity, Bachman describes the

'interactional/ability' approach to defining test authenticity as focusing on "... the *distinguishing characteristic* of communicative language use - the interaction between the language user, the context and the discourse." (Bachman, 1990: 302, original emphasis) This suggests that while interaction is indeed a key feature of communication, it is not an individual skill.

Others have gone further, arguing that interaction is 'co-constructed' by all the participants in an event (He and Young in Young and He, 1998). Summarising the work of earlier researchers with children in testing situations, Lazaraton concluded

they found that test results are really collaborative productions: ... the tester is more than just a conduit for questions, and test 'performance' is really a collaborative achievement. While they do not claim that interactional processes distort the test scores, they see the interviewer as more or less implicated in students' performance, because the assessment process is by nature coproduced. (Lazaraton, 1996: 155)

At one level, this seems a commonplace observation; 'interaction' is something that goes on between two or more people, who must therefore all be contributing something towards it. The practical difficulty it raises for test validation is whether it is either possible or fair to attempt a judgement on one person's interactive ability when at least as much is contributed by the other party or parties.

Another practical concern for the measurement of validity is whether interaction should be considered as a variable task by task or whether it is rather a higher-order factor that can only be treated at the level of the whole discourse. Canale and Swain's use of the superordinate term 'strategic competence' to include interaction skills (section 2.9.2) implies that it is something that extends over the longer discourse. In the design of the Five Star test, the term 'interaction' was used for strategic competence 'defined as coping strategies extending over pragmatic and sociolinguistic competence' with the suggestion

that "It may well be that this encompasses linguistic behaviour which has much greater stability than is normally intended by this aspect of SLA [second language acquisition] models" (Pollard, 1994: 40). The manual sequencing of the algorithm in some cases consciously makes possible the development of related themes over a sequence of several tasks (Pollard 1998a, see example in section 4.2), making the test more like a conversation and less like an interview (van Lier, 1989).

Concern with the centrality of interaction to oral performance testing and with its impact on validity is relatively recent:

There is no theory of [test] method to explain how particular aspects of method affect discourse and how those discourse differences are then reflected in test scores ...Nor is there a developed explanation of how rater and examinee characteristics interact with one another and with discourse characteristics to yield ratings, or how tasks relate to well-functioning scales (Upshur and Turner, 1999)

The original research design leading to the critical review of the Five Star test was faced with an algorithm underpinning the pilot test that allocated a participant's scores across six skills, listening, speaking, reading, writing, study skills and interaction. After some consideration, interaction was not used in the skills allocation rounds of the data collection for the reasons discussed above and does not figure in the expert panel data set:

The reasons for excluding the Interaction skill ... do not suggest a lack of importance attached to it, but quite the reverse, as the Five Star test can be seen to be centrally as a test of direct interaction between interlocutor and participant. However, interaction was felt a) to be much harder to define than the other core skills, reducing the likelihood of achieving a consensus; b) necessarily to overlap substantially with listening and speaking, where the methodology theoretically demands independence between skills; and c) to attach primarily to the event (the participants and the context of a particular test administration) rather than to the test item alone. (Underhill, 1997:3)

Lazaraton's own analysis of CASE (Cambridge Assessment of Spoken English) interviews found that interlocutors employed a range of verbal support strategies that

were typical of 'ordinary' conversation. The positive side of this was that it added authenticity, and hence some communicative validity, to the test process; but

on the other hand, the interlocutor support that occurred was not consistent, and this raises questions about its impact on candidate language use, and on the rating of that language (Lazaraton, 1996: 166).

What she recommends therefore is closer attention to the training of interlocutors, looking in particular at the types of support and behaviour that are appropriate to the role of the interlocutor. Where the market for a test is well-defined culturally, one might expect the interlocutor training to have specific regard for locally-sensitive judgements of appropriateness.

7.2.3 Implications of the communicative approach 3: task-based structure

There are both theoretical and practical reasons for adopting a task-based approach. At the level of principle, the communicative approach entails a task-based methodology as a consequence of construct, content and face validity. This theoretical underpinning for a task-based approach overlaps with authenticity, and together they reflect the fundamental communicative axiom that language is used to do things, and is therefore to be taught and tested as a means to an end, rather than as a body of knowledge in itself. One way to reproduce the functional use of language is to identify some of the things that we do with it in the real world, and to reflect these in a convenient unit, namely a 'task', in the classroom and testing context. This begs the question as to how this selection is to be made, unless a direct work-sampling approach is both valid and feasible; this is sometimes the case in teaching and testing for specific purposes, such as

air traffic controllers, waiters, or students preparing for a course of study with determinable topics and forms of interaction.

The approach to interaction as a core construct of strategic competence in the Five Star test draws support from the extension and development of themes over several tasks in a test event, whether this topic nomination occurs by preconceived design of task (Pollard, 1998a) or by spontaneous initiation by the candidate or the interviewer.

At a practical level, empirical validation tools work better with some kind of unitised structure, consisting of a sequence of more or less discrete events, rather than an undifferentiated whole, such as an unstructured oral interview. In this latter case, validity might be claimed for an open-ended and wide ranging discussion between candidate and interlocutor on other grounds such as authenticity, interactiveness, topicality and individualisation. However, the inability to identify distinct sections within the whole would make such a test very difficult to analyse internally. It would certainly not be possible to use IRT, or indeed to ask an expert panel's opinion about it, except in very general terms.

At the other extreme, traditional pen and paper tests consist of very large numbers of completely unrelated items, and the more such items a test contains the easier it is to establish and refine test reliability by statistical means. An oral equivalent might be a 'question-and-answer' test, where the interlocutor asks a series of fully- or semi-scripted questions, which the candidate answers and in so doing accumulates a score. The unrelatedness of the questions, although an advantage in meeting the assumptions

underlying some statistical procedures, directly violates the communicative criteria of meaningful and authentic interaction.

A task-based approach might therefore be seen to be a practical compromise lying somewhere between these two extremes. It provides a modular structure in which the pieces can be fitted together in different permutations to contribute to an overall picture, yet each piece is large enough in itself to claim a degree of real-life authenticity. The modular nature is ideally suited to an adaptive algorithm, where the sequence of tasks is neither completely pre-determined nor totally random. Item banking is one of the most common uses of item-response theory (IRT), as it possible to 'characterise the items in a way that is stable from one sub-population of testees to another' and so allow tests to be constructed from different items yet remain comparable in difficulty and statistical equivalence (Baker, 1997: 50).

Two other practical advantages accrue from a task-based approach. Firstly, it allows the trialling of new tasks in a process of continual development. Pilot tasks can be integrated and tried out in authentic test contexts, but their scores can be separately analysed and not contribute to the whole test score or profile. Secondly, different tasks may also be consciously selected to vary the skills being tested; to focus on different proficiency bands, in the belief that some task types generically require a higher or lower minimum level proficiency band than others (although Fulcher, 1996, found no evidence to support this for oral tasks); or as vehicles for different rating scales, for example, focusing on range of vocabulary or accuracy in one task and fluency or size of utterance in another.

Unfortunately, this still doesn't help define what a task is, other than by duration - it is longer than a single question-and-answer but shorter than a conversation. The definitions quoted in chapter three suggest among the possible criteria an emphasis on exchanging meanings rather than just producing language forms, and the setting of an identifiable goal, the achievement of which might not be predictable; in other words, unlike a single test item, there is not a single correct answer.

7.2.4 Implications of the communicative approach 4: individualisation

Performance, in its most general sense, allows the possibility of individual variation in behaviour. Artistic or sporting performance may be judged solely in terms of comparison of better/worse or faster/slower judgements, but typically the description of performance is richer than that. A performance test that aims to elicit a demonstration of the use of language skills must also accommodate differences in behaviour.

Section 2.14 drew a distinction between individualisation deriving from the communicative approach (personalised to an individual) and individualisation deriving from social factors such as a reaction against norm-based testing (personalised to a group or distinct population). The prominence given to interaction, and the desirability of an element of unpredictability (Weir, 1990, cited in 2.9.2), is likely to add a further source of 'random' individualisation:

The sociological 'rules' for testing interactions are always embedded in a particular context, and the processes of interaction in these contexts will vary from subject to subject, even if the outcomes (i.e. scores) are equal (Lazaraton, 1996: 155).

The implication is that in other cases the outcomes will not be equal, and that unpredictable interaction as an individualising design feature will lead to a lack of comparability between test events.

Some of the features of the communicative approach already discussed necessarily imply variation between test events and between language samples generated from different candidates. Test techniques that draw on more than a knowledge of language structure yet yield single correct answers only might be seen as the 'holy grail' of language testing, and a great deal of effort and ingenuity has gone into the development and justification of the kind of pragmatic and pseudo-communicative tests described in chapter three. To generate standardised or easily-comparable test outputs requires a uniform input or set of instructions, and a conscious disregard for whatever the candidate may say or do in response.

Thus, a live question-and-answer sequence is a direct test, and may superficially appear to be individualised, but if it is pre-scripted and the candidate's utterances actually have no influence on the interlocutor's subsequent behaviour, then such test procedures fail to match the criteria of lifelikeness and interaction, however convenient they may be from the point of view of marking. Interaction, in particular, implies individualisation. Tasks that invite the candidate's opinions or personal experience are more likely to lead to individualised interaction than those which ask for the display of general knowledge about the local context or state of the world.

Although adaptivity and individualisation overlap, it is important to distinguish between them. In its specific sense, adaptivity is a technical feature of a test that responds only to

the candidate's level of proficiency, irrespective of the communication involved. Individualisation is a communicative feature of the test event that entails responsiveness to the meanings being conveyed, whether or not this responsiveness is linked to proficiency level. Three examples can be used to illustrate this distinction: one that is adaptive only, one that is individualised only, and one that is both.

Example 1: a computerised multiple choice test of grammar can be adaptive, but need not be interactively individualised. The candidate sits at a screen and is presented with a sequence of items according to a pre-determined algorithm which reacts to earlier responses, and with a large enough item bank to draw on one candidate's test sequence may actually be unique, but this is a focus on language form only. There is no individualisation to the content of the candidate's responses or his or her ability to communicate meaning.

Example 2: a traditional live oral interview focusing on content and meaning is often individualised, as discussed in chapter three, by the interlocutor steering the interaction to respond to what the candidate says. The interlocutor does this by picking up messages about which topics the candidate feels more comfortable with and is able to expand on, thus generating a larger language sample on which to base a reliable assessment. The interlocutor may have a choice of tasks to present to the candidate, but this choice may only be alternative tasks of more or less comparable difficulty, rather than a conscious decision to present an easier or harder task in the light of previous behaviour. In such cases, the test is individualised to the candidate's expressions of meaning but may not be adaptive to the candidate's level. This is typically the case for

oral examinations where the approximate level of performance is defined in advance, and the outcome is a pass/fail judgement, rather than a score on a wide-ranging scale.

Example 3: a computer-based but interlocutor mediated test can combine both of these. The computer algorithm provides the formal basis for adaptivity and the selection of subsequent tasks, while the human interlocutor individualises the interaction with the candidate during each task. Experienced oral interlocutors may be able to combine both forms of responsiveness on their own, as claimed for the FSI (American Foreign Service Interview) test in chapter three, but the adaptivity is likely to be based on subjective intuition. The advantage of the computer-based method is that it permits individualisation of content while retaining the immediate objective analysis of task-based data for level.

Ultimately, the distinction between adaptivity and individualisation may be hard to maintain. Particularly in live oral interviews where interlocutor has the discretion to vary the direction of the interaction, or the candidate has some choice over the task to be completed, it would in practice be very difficult to separate which choices were influenced by considerations of proficiency level and which were not.

7.2.5 Implications of the communicative approach 5: topicality

The communicative function of language is linked to the exchange of meanings between certain participants at a certain time and in a certain place. While these contextual features of the situation do not dictate the topic of language use, because we can in

principle talk about anything anywhere, they do in fact have a strong influence on it, and the communicative approach therefore requires a consideration of topicality as well as topic.

The selection of possible topic bias in testing materials is an issue which affects all tests, not just communicative ones. In brief, the ideal topic is one on which all candidates are equally knowledgeable, so that it will favour no candidate for non-linguistic knowledge of the world. In practice, this is extremely difficult to do, particularly when at the same time the test writer is trying to select a topic that will be of some interest to test-takers, and only by deliberately choosing a topic that is extremely general and bland can this possibility be avoided. Where the bias is systematic for or against an identifiable population, is it sometime called 'group bias' or 'differential item functioning' (e.g. Zumbo, 1999; Henning in De Jong and Stevenson, 1990) Of particular concern in both UK and USA has been the possibility that 'high stakes' academic admissions tests used for applicants from all nationalities and backgrounds might exhibit differential item functioning, for example, when tests of academic English based on one discipline are used for applicants from another. Not surprisingly, studies often show that students scored better on reading tests based on texts from their own discipline, but the evidence is far from unanimous (Henning in De Jong and Stevenson, 1990; Fulcher, 1996).

One way out of this has traditionally been to offer test candidates a limited choice of topics, for a written essay or oral presentation for example, or more recently with the development of continuous assessment methodologies to allow them to select and

develop their own topic. A performance test in which a candidate has no choice, or no possibility of negotiation, of topic could be seen as potentially biased.

The communicative requirements of meaningfulness and authenticity add in addition an extra requirement, that the topic should be relevant to the participants. Three possible aspects of relevance can be distinguished here: temporal, geographical and cultural relevance.

- a) Temporal relevance reflects the fact that much of our real-life conversation is concerned with current or recent affairs. This is necessarily ephemeral; something that is 'newsworthy' by virtue of its currency today may be fading from public concern next week, and forgotten the week after. This may therefore be a problem where test topics must be chosen a long way in advance, and may conflict with the need for programming or exhaustive trialling. It may be a particular problem for tests aimed at teenagers or young adults, whose perception of what topics are 'in' may change even faster than other groups' perceptions.
- b) Geographical relevance reflects the local, national or regional nature of the events that concern us most. This is not ephemeral, and so potential test topics have a longer lifespan, but it may restrict the transferability of a test. Major events in one country or region may be viewed with indifference in neighbouring countries or regions. This may overlap with cultural relevance where there is a strong sense national identity.
- c) Cultural relevance reflects differences between societies, or different groups within a society, about what are common or permissible topics. In some cultures, any reference to politics or religion is at best unwise and at worst may cause offence. Football and cars may be ubiquitously suitable topics for young men, but not for

women, especially in countries where there are restrictions on their public behaviour.

As with selection of topic in general one partial solution to these topicality constraints is to provide some flexibility, within a task of a particular format, for choice by interlocutor or candidate or negotiation between them. Within a task that invites a candidate to express an opinion, for example, the interlocutor can substitute an appropriate and relevant topic. There is however an assumption here that there is no interaction between topic and task type, which may not in fact be justified.

Overall, relevance of these different types is essential for communicative validity, but this immediately adds restrictions of time, place and culture to any validation evidence that is gained. In theory, any transfer of the test to a significantly different location, any culturally different target group, or any great lapse of time since the validation evidence was gained would require fresh evidence to be collected. In practice, it is often difficult to judge when such a transfer has in fact taken place, and the only practicable policy is therefore to collect and analyse such evidence on a continuing basis.

7.2.6 Implications of the communicative approach 5: authenticity

Many of the issues that are raised by the communicative requirement of authenticity have been discussed in the previous sections, as they overlap with concerns of interaction, task-based orientation, interaction and topicality. Indeed, part of the difficulty of discussing authenticity, illustrated by the different views in the literature

cited in 2.14, is to determine its status as a construct in relation to interaction, for example. Is authenticity a positive attribute that flows from a test being genuinely interactive, or is authenticity a superordinate construct, with interaction, topicality etc as variables that operationalise authenticity in different ways?

The Bachman and Palmer (1996) checklist offers an attractive approach to determining authenticity; it is simply the extent to which test tasks match the characteristics of specific real-life or target language use situations. The deterministic model would have us believe that this is a straightforward process, but even if this was the case, it does not allow for the embedding of professional language use in general conversational language use: "A specific skills approach to oral proficiency therefore does not obviate the need for an evaluation of general conversational ability" (van Lier, 1989: 500)

The reference to Searle's speech acts in section 2.14 as another approach to authenticity provides a possible framework for examining the authenticity/interaction of individual tasks in a test. The typical 'exam question' which requires display of knowledge can be compared to a pen-and-paper test; the examiner already knows the answer, and doesn't genuinely want the information, but does want to know if the candidate knows. But with a task-based approach to a performance test, there are many possible questions to which the speaker does not know the answer, so that one of Searle's conditions for a 'real question' is easily fulfilled.

The second condition, that of sincerity, requires not only that the speaker does not know the answer but that he or she genuinely wants the information. The extent to which this can be met depends both on the structure of the tasks and on the behaviour of the

individual interlocutor, and here again authenticity overlaps with interaction. A question-and-answer routine in which the interlocutor asks a series of personalised questions, the answers to which will be different for each a candidate, will satisfy the first condition but not the second if the interlocutor does not need to follow up or makes no attempt to build on the candidate's responses.

However, if the interlocutor's directive is to create an interaction then it is in his or her interest to create and maintain the interaction. According to one view quoted by Spolsky

there can be no test that is at the same time authentic or natural language use; however hard the tester might try to disguise his purpose, it is not to engage in genuine conversation with the candidate, or to permit him/her to express his/her ideas or feelings, but rather to find out something about the candidate in order to classify, reward, or punish him/her (Spolsky, 1990: 11)

but the suggestion here is that in fact genuine interaction is possible if both interlocutor and candidate feel it is in their mutual interest.

To determine authenticity, by this measure, we would look for tasks that are more likely to engender lifelike interaction, and as a measure of this, transcriptions or records of test events that show the interaction being extended beyond the minimum necessary for transfer of the information requested. We might also look for task rubrics that direct interlocutors to create and maintain such interaction. We might start with video tests of interlocutors who seem to be able to create and maintain genuine interaction more easily than others, and proceed to transcription or other forms of analysis to identify the key ingredients or behaviours.

However, the authenticity of such extended interaction again encourages a diversity of different directions that the conversation can take, and so a greater likelihood of

variability in behaviour between interlocutors. While variation in speech behaviour between individuals in a similar context might be seen as a positive feature of lifelikeness, in a testing context it is a potential source of error undermining the comparability of samples elicited from candidates and thus the reliability of the test.

7.3 Summary

The theoretical contrasts considered in the first section provide ample justification for the inclusion of as wide a range as possible of types of data in validation studies, but in many cases they raise questions that can only be answered with respect to a particular test on the ground in its local context. While there is considerable agreement about the desirable features of a communicative approach to testing there is similarly no consensus about how to integrate them in a coherent model; this is in essence Davies's 'countable shotgun' approach quoted in chapter three above.

Future research into the nature of the testing process may uncover a more explicit combinatorial analysis of theoretical issues and communicative test features. In the meantime, the best way forward for the practical purposes of test construction may be to accept that the current consensus about the meaning of the communicative approach to language testing reflects the current paradigm for test validation; they represent lists of desirable features, which can only be realised, prioritised and evaluated in the light of the local context of each specific test. The next chapter attempts to combine the checklist approach with the typical development sequence of a new test in a tightly-integrated model for continuous validation.

Chapter 8 Derivation of theoretical model for continuous validation

- 8.0 Introduction
- 8.1 Preliminaries
 - 8.1.1 Purpose of test
 - 8.1.2 Development plan, including commercial and strategic considerations
 - 8.1.3 Description of target language use (TLU) domain
 - 8.1.4 Identification of central construct(s)
 - 8.1.5 Establish programme evaluation
 - 8.1.6 Establish expert panel
- 8.2 Checklist of test characteristics required for the validation model
- 8.3 The validation model
- 8.4 Further description of model components
 - 8.4.1 Component 1: programme preliminaries
 - 8.4.2 Components 2 and 3 : pilot test cycle and main test cycle
 - 8.4.3 Component 4: Subsidiary test cycles
 - 8.4.4 Component 5: programme evaluation
- 8.5 Summary and imitations of the model

8.0 Introduction

This chapter aims to apply the critical analysis of the previous chapter to the derivation of a theoretical model for continuous validation, in pursuit of the aim to design a model appropriate for an adaptive and communicative language test that allows for the validation to become a recurrent process as the test evolves.

The checklist of characteristics required for the model is extrapolated from the literature review and critical thinking on test validation and communicative methodology in previous chapters. These requirements are considered individually in sections 8.1 and 8.2 and how they fit together to form the components of the model in 8.3. Although the model is primarily concerned with the validation of the test itself, this process must be seen as part of the wider real-world activities which have implications for the applicability and conduct of this approach to validation, but are not part of the validation process itself.

The central construct under measurement is included as a component of the checklist to be defined, but in reality the relationship between the construct and the overall approach to validation is iterative. Subsequent modifications to the principles of the communicative approach may also modify the list of characteristics, and a complete paradigm shift in language teaching and testing that replaced the communicative approach might fundamentally rewrite the checklist and the model.

8.1 Preliminaries

The model that is being presented here focuses specifically on the validation of a test. It is not an *ab initio* model for test design and development (such as is offered in Bachman and Palmer, 1996) and therefore makes certain assumptions about preliminary activities that have already been undertaken and whose outputs would normally be taken as a 'given' for test validation. The APA Standards require that "the purpose of the test,

definition of the domain, and the test specifications should be stated clearly so that judgements can be made about the appropriateness of the defined domain for the stated purposes(s) of the test..." (APA, 1999: 43). For the sake of completeness these preliminaries are summarised in Table 35 and then discussed individually. In the normal course of events they will require routine review, like all the other components of the model.

Table 35 Summary of preliminaries to test validation model

Test characteristic	Processes and possible sources of data	Outputs from process
8.1.1 Purpose of test and identification of interested parties	Origin of test Survey of stakeholders	Statement of purpose Stakeholder analysis identifying different interests and extent of involvement in operation and results of test
8.1.2 Development objectives, including commercial and strategic considerations; design of the algorithm	Formulation of business and development plans Origin of test Survey of stakeholders	Business and development plans
8.1.3 Target language use	Analysis of target language use; ideally, based on data from direct observation of language use in context	Description of target language use (TLU) domain
8.1.4 Central language construct(s) and test specification	Identification and justification of central construct and test specification; reference to statement of purpose and stakeholder analysis Costing process to identify costs of different test options	Definition of central construct(s) and test specifications and test specification Cost/benefit analysis
8.1.5 Programme evaluation framework	Formative / summative programme evaluation Reference to end-users and other stakeholders	Programme evaluation framework 'Usefulness' rating
8.1.6 Establish expert panel	Recruit, select, train, moderate and establish modus operandi for expert panel	Statement of panel policies and procedures

8.1.1 Purpose of test and identification of interested parties

The original motivation for the development of the test will provide an initial statement of the purpose(s) for which the test will be used and the likely use that will be made of the results. The statement of purpose may imply a definition of the central construct under measurement - e.g. communicative competence - which needs to be made explicit at a later stage, 8.1.4 below. The starting point can be phrased as a question of beliefs about language proficiency assessment which then informs the fundamental construct (8.1.4) and overall model of validity (8.1.6): 'Do you / your stakeholders believe that what is measured by tests of communicative language performance is sufficiently important, and sufficiently different from pen-and-paper tests of language knowledge, to justify the additional resources required for development and maintenance?'

A more detailed analysis of all the stakeholders will pick out the different interests and likely degrees of involvement of interested parties. This process may be sensitive. The stakeholders might include :

- end-user institutions (academic, governmental or commercial) requiring test results
- staff of the host institution, both those directly and indirectly connected to the testing programme
- test takers
- other governmental or quasi-governmental organisations with a responsibility for this area of education, training or employment
- institutions preparing or entering candidates for the test or from which candidates are likely to apply
- institutions involved in funding the test

The recurrent aspect of the stakeholder analysis may be termed 'stakeholder feedback' and this can play an important continuing role as a subsidiary activity to the main test cycle, where it specifically monitors reactions to test events and outcomes. This could take the form of

- feedback from test participants on test events and on test outcomes
- the perceptions of test administrators, interlocutors and assessors on how well the task/test performed
- feedback from end-users (employers, admissions tutors) - also feeds into external validity

8.1.2 Development objectives, including commercial and strategic considerations; design of the algorithm

This is likely to be a commercial activity, taking as its starting point the original motivation for test development. The financial plan must allow for the continuing validation of the test and evaluation of the overall programme, costed on the kind of activities proposed in the model, as an integral and recurring expenditure. It must also take into account the likely costs of the type of test event to be routinely used, allowing for example a higher cost for a direct oral test than for pen-and-paper written test. Since this may not become clear until later in the elaboration of the test format, there needs to be an iterative link back to this plan from later stages.

The development plan will among other factors consider issues of human and technical resources. Assuming resources are not infinite, choices will have to be made between ideal circumstances and real-world constraints, e.g. in the frequency of cycle, number of tasks in the task bank or the size of pilot sample. The development plan may therefore allow for a more restricted implementation of the model to begin with, gradually being expanded in line with the growth of the test. Other external factors affecting the operation of the model may include the local academic or recruitment calendar and the technical platform used for the test. A requirement for external programming and production, e.g. of a CD-ROM, implies fewer, larger-scale revisions of the computer-based test than would be possible with in-house capability of revising a disk-based test.

An important theoretical decision to be taken early in the development plan concerns the nature of the algorithm, the process that determines the sequence of test tasks to be presented in each test administration. A large-scale test will typically rely on automatic selection of a task at random from amongst those at the appropriate difficulty level in the task bank. In other words, after the first task has been administered, subsequent tasks are selected in the light of previous responses, to fine-tune the distribution of tasks administered to the proficiency level of the candidate. Where multiple skills are also being tested, then the algorithm may also consider the skills profile of tasks in the selection process. This greatly increases the size of the development task to create, pilot and maintain a large pool of current tasks. Alternatively, the algorithm can be designed manually, creating the links between tasks which effectively determine all the possible routes through the test. This is much easier to set up initially, and allows routes to follow a theme or sequence of topics over more than one task, but will require periodic

manual revision to reflect new tasks or revisions to task data such as difficulty or skills tested.

8.1.3 Target language use

The statement of purpose in 8.2.1 implies a target audience and/or contexts of language use for the test. This needs to be made explicit and defined sufficiently to allow an analysis of the actual language being used. There are different possible models for this. Munby's detailed communicative needs processor (Munby, 1978) identifies needs exclusively in the real-world contexts through lists of skills and sub-skills, communicative events and language functions. Weir's (1990) list of parameters is based on Munby and is specifically test-based but much briefer. Alderson et al. (1995) surveyed the approach to test specifications used by different examining boards. Bachman and Palmer's (1996) framework of language task characteristics is deliberately ambiguous, embracing both real world and test events. In all cases, the output is a description of the target language use. Other frameworks for content specification are designed for ESP (English for Specific Purposes) contexts rather than general language use, such as Carroll (1980) and Hutchinson and Waters (1987).

There is a range of practical, procedural and linguistic problems. Access to target language users may be restricted, and a process of observation and analysis may affect the naturalness of the language used. Ethical and practical considerations may make recordings of spoken samples difficult. There are unresolved questions about determinability and the extent to which language used in even quite specialised contexts

can be considered different from everyday language, except in trivial and obvious ways such as the use of specialised terminology and professional jargon.

8.1.4 Identification of central language construct(s) and test specification

The test specification should be a complete blueprint for the construction of the test as a whole which acts as a template within which the description of target language use fits to generate individual items. In the ideal world, the specification draws crucially on the definition of constructs, which in turn are based on an underlying theory of language learning and use.

However, following on from identification of purpose and interests in 8.2.1, in the real world the central construct under measurement may already be implicit or explicit. At one extreme, it may be an externally-derived given, originating for example from ministry of education specifications. At the other extreme, stakeholders may take an undifferentiated view of language proficiency, and may simply regard this as a minor detail to be delegated to the experts. The risk in the latter case is that the extra costs of performance tests requiring the elicitation and judgemental rating of samples of reading and writing may require justification. In either case, the pedagogical approach for teaching and testing that is locally dominant may not be appropriate to meet the identified test purpose, and this divergence may be confirmed by the stakeholder analysis.

In such cases, a cost/benefit analysis can be used to identify the different possible test formats and their likely costs. In some circumstances, a consensus might be established by a survey of end-users; the fundamental question would be 'Do you / your stakeholders believe that what is measured by communicative performance tests is sufficiently important, and sufficiently different from pen-and-paper tests, to justify the additional resources required for development and maintenance?'

The central construct will itself be hypothecated on an underlying theory of language, which should be made explicit rather than left as a tacit assumption; a few authorities on testing (Oller, 1979, and Bachman, 1990, for example) provide a substantial foundation of language theory on which to build tests, but many do not. Without a theoretical underpinning, establishing and operationalising the central construct will be more difficult. Where there is a group of linguistically-skilled personnel, a Delphi-type exercise may be used to establish consensus on operational definitions (e.g. see section 5.1.1). With the exception of Munby (1978), the authorities cited for the description of target language use in the previous section also give procedures for test and item specification.

8.1.5 Establish programme evaluation

Although test validation has an external perspective through consideration of consequential validity, for example, it is essentially an internally-oriented activity, focusing on the content, structure and behaviour of the test. Programme evaluation in contrast is an externally-oriented activity that embraces test validation among its

sources of data but has a broader remit, ultimately answering the question of whether the test successfully achieves its purpose or not within the parameters of the business and development plan. Language testing is never done in isolation, and programme evaluation allows it to be located in the wider local context. The diagrammatic model presented in Figure 8 in section 8.3 below therefore has 'internal (test) orientation' vs. 'external (programme) orientation' as one axis for describing the different validation activities.

There is some potential overlap between programme evaluation and specific forms of validity data, e.g. comparison with external criterion tests and consequential validity. The context will determine how distinct and detailed the programme evaluation should be. A simple version would take Bachman and Palmer's concept of 'usefulness' as a superordinate term including other sources of data such as practicality and impact as well as validity. A more formal version could be adapted from the literature on language programme evaluation, and a possible model for this is discussed in section 8.4 below.

As in other fields, there is a tension between positivistic/quantitative and naturalistic/qualitative approaches. In language programme evaluation specifically, the quantitative approach has favoured large scale experimental designs comparing two or more teaching methodologies and leading to a summative judgement. In an overview of language programme evaluation projects, Beretta considers 33 such studies between 1963 and 1985, including such famous names as the Pennsylvania and Colorado projects, in some cases involving thousands of participants over a period of several year (Alderson and Beretta, 1992: 10). The most recent of these, the Bangalore project,

compared communicative and grammar-based approaches in schools over a period of four years.

However, there was a growing realisation in the 1980s that

without a description and clear understanding of the process (i.e. what actually happened in the program as well as what happened in the control or comparison situation) there would be many plausible explanations for the outcomes of product evaluation (Lynch, 1996: 32)

and the emphasis swung towards process-oriented, naturalistic approaches favouring rich sources of data to produce formative evaluation.

Most recent models have contained similar components in the form of a checklist or series of steps. Alderson and Beretta (1992) propose a basic model that is taken up in section 8.4.4 below. Other models allow for specific complex factors to be analysed in more detail, such as stakeholders (Nunan, 1992), staff motivation (Mackay, 1993; Wellesley, 1993) and context (Lynch, 1996).

8.1.6 Establish expert panel

For the purposes of this model, the role of the expert panel is seen as external to the regular development and delivery of the test, which is carried out by staff collectively referred to as test developers or administrators. The extent of routine involvement by the panel may vary according to context, but in principle they are independent.

The expert panel operates at two levels of scrutiny and at both the pilot and main cycle stages. The two levels of scrutiny are the individual task and the overall test. In the pilot

cycle, the panel scrutinise draft tasks proposed by the test developers, using the characteristics 8.2.2 - 8.2.7 below as the primary criteria for judgement.

In the main cycle, they again scrutinise proposals for new tasks and also review existing tasks, as a routine 'task banking' activity and in the light of performance data generated by test administrations. Such review might lead to changes to any aspect of the task, including rubric, content, rating scales used, location in algorithm and skills allocation. In the main cycle, they also comment on the overall test level, but this task is made more difficult by the adaptive nature of the test. Samples of tests, for example recorded on video, can be viewed by panel members and used for panel moderation as well as test review purposes.

Views in the literature differ about how the panel should be composed (see section 5.1.1) and the exact ambit and method of procedure will depend largely on the local circumstances. The most important general principles are external independence and recurrent activity:

- the panel should be independent of the test developers
- the panel should comment, criticise, advise, recommend on the development of tests and tasks but should not directly originate them
- like other components of the model, the panel's activity should be based on a recurrent cycle of activity
- the panel's role in the review process and its members' qualifications and background should be documented (APA, 1999)

In some circumstances, recruiting a panel which has the requisite degree of expertise in the field, is locally knowledgeable, and sufficiently independent of the test development

programme may be impossible. However, the panel may not need to meet physically, and indeed need not be located in one country or area. Participation by electronic media would not only make possible the rapid collection of comments and discussion of tasks/tests but would also facilitate the 'anonymous consensus' type of panel exercise and allow a wider pool of panelists to draw on.

8.2 Checklist of characteristics for the model

Table 36 presents a checklist that draws on the literature and discussion in previous chapters to identify the different characteristics of the model to be constructed in the next section. These characteristics are then discussed individually in sections 8.2.1 - 8.2.12.

Table 36 Checklist of test characteristics for the validation model

Test characteristic	Processes and possible sources of data	Outputs
8.2.1 The overall model of validity	<p>Consideration of diverse possible sources of data</p> <p>Reference to statement of purpose and stakeholder analysis</p> <p>Reference to definition of central construct</p> <p>Reference to end-users and other stakeholders</p>	<p>Statements of principle and operational policy on validity.</p> <p>List of feasible validation activities and potential sources of validity data</p>
8.2.2 Direct testing	Scrutiny of tasks by external panel and/or stakeholders against TLU	Qualitative evaluation; recommendations for improving tasks
8.2.3 Interaction	<p>Scrutiny of tasks by external panel and/or stakeholders against TLU</p> <p>Conversation / interaction analysis of actual test events</p> <p>Feedback from stakeholders</p>	<p>Qualitative evaluation</p> <p>Recommendations for improving tasks</p> <p>Review of interaction construct</p>
8.2.4 Task-based structure	<p>Scrutiny of tasks by external panel and/or stakeholders against TLU</p> <p>Management of 'task bank'</p>	<p>Qualitative evaluation; recommendations for improving tasks</p> <p>Procedures for operation of task-bank and task life-cycle</p>
8.2.7 Individualisation	<p>Design of algorithm, tracking of actual test paths</p> <p>Scrutiny of test records, e.g. video or audio recordings</p> <p>Interaction analysis</p>	<p>Recommendations for improving design of algorithm</p> <p>Recommendations for improving interlocutor training and moderation</p>
8.2.6 Topicality	<p>Scrutiny of tasks by external panel and/or stakeholders against TLU</p> <p>Task 'life-cycle' and other data from task bank</p> <p>Stakeholder feedback</p>	Qualitative evaluation; recommendations for withdrawing, replacing, improving tasks
8.2.7 Authenticity	Scrutiny of tasks by external panel and/or stakeholders against TLU	Qualitative evaluation; recommendations for improving tasks/test

Table 36 (continued) Checklist of test characteristics for the validation model

Test characteristic	Processes and possible sources of data	Outputs
8.2.8 Content validity	Scrutiny of tasks/test by external panel and/or stakeholders Comparison against TLU Comparison against other external tests IRT item-fit data	Qualitative evaluation; recommendations for improving tasks/test
8.2.9 Face validity	Scrutiny of tasks/test by external panel and/or stakeholders	Qualitative evaluation; recommendations for improving tasks/test
8.2.10 Empirical validity	IRT data and other task-bank information Expert panel Interaction analysis Stakeholder feedback Scores on external criterion tests	Qualitative evaluation; recommendations for improving tasks/test
8.2.11 Reliability	Recruitment of rating panel; training and moderation Scrutiny of data from test administrations (video, audio, paper-based) IRT analysis of rating data as a distinct test facet	Consistency of judgements Data about rater performance against specified standards
8.2.12 Consequential validity External validity	Panel judgements Concurrent and predictive criteria and judgements Stakeholder feedback Impact assessment	Qualitative evaluation; recommendations for improving tasks/test

8.2.1 The overall model of validity

This is an enabling activity that does not itself generate primary data but which is necessary to authenticate other validation activities in the model.

Overall, in the current paradigm, validity is seen as a unitary construct based on diverse types of evidence. A model based on this must therefore allow and incorporate collection of different possible types of data from diverse sources: empirical and naturalistic, pre-determined and emergent variables, at pre-determined stages and as a continuing process. The underlying approach to validity data collection may be characterised as principled eclecticism. In real-world contexts, it is not possible to be dogmatic about types of data and methods of combination, but rather one has to seek data where it is available in the local context and allow for rich types to emerge.

However, where the sources of validity and the ways they may combine are not determined in advance, it is all the more important that the programme evaluation activity (8.1.5) lies outside the model of validity in order to retain an overall view of the validation process. There is a risk otherwise that it may become purely opportunistic and 'anything goes'.

8.2.2 Direct testing

This characteristic, and the following five 8.2.3 - 8.2.7, form the immediate focus of concern of the panel at the 'task' level of scrutiny at both the pilot and main cycle stages. Although it may be useful to present them as five distinct criteria for evaluating tasks, in reality it may be difficult for the panel to make separate comments on such characteristics as interaction, topicality, authenticity and directness without overlapping.

Superficially, directness is a very easy criterion on its own; either a task/ test is direct, in that it requires face-to-face two-way communication between two or more participants, or it isn't. However, within the class of direct tests that meet this criterion, there are a range of factors to consider, such as

- the extent of separation of test administrator, interlocutor and assessor roles
- multiple-skill tasks and task types; are tasks that call for different skills such as reading and listening still delivered within a 'direct test' framework?
- specific behavioural descriptors for rating scales; do they accurately reflect direct target language use?

Traditionally, direct tests such as oral interviews have been organised as separate events from pen-and-paper tests, even where they form part of larger test batteries (e.g. UCLES, current). Some direct tests have implicitly or explicitly included scope for adaptivity, but computer-based tests, whether adaptive or not, have not had any direct component. Indeed, it has been the automation of testing by computers, and therefore the potential cost saving, that has been one of its major attractions. It is the hybrid test combining both face-to-face and computer-based characteristics that is the particular focus of this research.

8.2.3 Interaction

As was seen in the discussion in 7.2.2, this is one of the most intractable constructs in language testing. Again, it may seem superficially easy for the panel to determine whether interaction is taking place or likely to take place, but very much more difficult to quantify how far it matches interaction in the target language use. More than any other test characteristic, interaction is dependent on the participants and the context of each test event, and a prospective surmise of the possibilities for interaction, based only on a scrutiny of a task outline on paper, may be wildly inaccurate.

The model should therefore include some form of conversation / interaction analysis as one source of validity data, and this is built into the model as one of the subsidiary test cycles. As well as generating data for improving individual tasks, this will also provide information about the role of interaction in the test, and ultimately allow an informed review of how it is operationalised within the validity model. Practical questions such as 'can interaction be measured as a distinct language skill?' can then be addressed.

8.2.4 Task-based structure

The preferred test model is based on tasks for the theoretical and practical reasons discussed in 7.2.3. While the literature suggests that 15-20 tasks is sufficient to reach an accurate score, the direct nature of the test may allow the interlocutor to vary task length within the existing structure so that he or she can be personally confident of the score being awarded.

In the pilot cycle, the panel consider draft test tasks proposed by test developers. In the main test cycle, they consider new task proposals, review existing tasks in the light of data from task bank records, and also comment on how the algorithm combines tasks to make an overall test.

Task banking is directly analogous to 'item banking' (Jones, 1991; Henning, 1987) and a major subsidiary activity to the main test cycle is the management of the 'task bank'. This is in essence a database containing records about each task with the following data:

- specific aim of task in terms of TLU, however analysed
- development history of task, with developers' and panelists' comments
- results of trialling and subsequent amendments (the reporting of results for pilot item results should be separately reported and excluded from collective IRT analysis for the test as a whole)
- results of main cycle administration in the form of stakeholder feedback; IRT data on task difficulty and fit
- evidence of skill or skills tested, from different sources

- results of continuing panel review and recommendations for extending or terminating task-life
- information about the task from interaction analysis
- frequency and pattern of use by the algorithm

8.2.5 Individualisation

Computer-based adaptive tests are ideally suited to individualising test events while retaining the systematic collection and analysis of data that allows comparability of performance.

There are three possible sources of individualisation. The first stems from the algorithm, the computer programme that underlies the selection and sequencing of task presented to each candidate. It is designed to respond to candidate performance by selecting tasks of an appropriate difficulty level, and it may at the same time ensure an appropriate balance of tasks presented by skill. Although this responsiveness is purely in response to candidate proficiency rather than interest or personality, it quickly generates different test events and with a large enough task bank will generate unique tests. Being computer-driven, this type of individualisation is entirely predictable and supposedly objective; in fact, a lot of the data that inform the algorithm and the decisions made as a result will always be subjective.

The second source of individualisation is the nature of the task itself. A task that asks candidates about an aspect of local culture or regional geography can be marked by a rating system that rewards language performance rather than display of general

knowledge, but will tend to generate standard responses rather than discriminating between candidates. However, a task that asks for a candidate's personal experience of that local culture or regional geography has the potential to be much more highly individualised.

The third source of individualisation is the interaction between candidate(s) and interlocutor. As noted above, the interlocutor may use his/her discretion to curtail or extend the interaction in order to obtain a satisfactory sample of language on which to base a rating decision; in theory, the candidate(s) may also choose to control the direction of the conversation and introduce new topics, but in practice the implicit power imbalance makes this less likely. However, more open-ended tasks certainly allow the candidate to introduce personal information and opinions to which the interlocutor may respond, and this effectively makes each test event unique. Where such interaction is not specified by the task rubric, it becomes almost a random occurrence, and raises the methodological difficulty that the more individualised a test becomes the harder it is to compare test events and thus to justify comparability of scores.

The extent of this variation can be monitored qualitatively through interaction analysis and quantitatively by timing records of the duration of each task, either manually or automatically through the computer-based test record. Evidence should be fed back into the rater training cycle on the effectiveness of different interaction strategies.

This scope for individualisation in principle provides evidence for strong construct validity for adaptive testing within the communicative paradigm; but it also raises questions about the resource requirements of communicative testing, and requires

evaluative judgements to be made about how large a scale of individualisation of each test event the overall programme can routinely afford.

8.2.6 Topicality

The communicative need for topicality discussed in 7.2.5 imposes a requirement for tasks that are geographically and culturally relevant as well as up-to-date. The task-banking management referred to in 8.2.4 above must incorporate a time-limitedness for each task with an estimate of its likely expiry date. This 'lifespan' is neither determinable in advance nor the same for all tasks; in any culture, some issues remain matters of real public debate indefinitely, while others do not. In general, medium-term trends may be more suitable as the basis for discursive tasks than short-term news headlines, on the same basis that magazine or periodical articles are a better source of classroom and testing material than newspaper articles.

It is one of the major routine activities of the expert panel to review tasks to assess whether they are still sufficiently topical or not. Recommendations by the panel may be to withdraw a task altogether, to amend it to bring it up-to-date or to leave it as it is for a further period. As well as evidence of the tasks and task performance data, demographic changes to the candidate population or variations in the TLU identified may influence recommendations about retention.

8.2.7 Authenticity

As suggested in 8.2.2 above, because of the difficulty of uniquely defining authenticity, it might form part of a 'bundle' of criteria that the panel are invited to consider each task against. As well as considering the tasks themselves, the panel should also review the specific behavioural descriptors used in any rating scales, to decide whether the rating scale descriptors match the TLU, for example, by reflecting the factors that determine success or failure of communication in real life.

8.2.8 Content validity

The content validity is established from diverse sources, identified by the overall validation model at 8.2.1 above, but what the model cannot do is predict how these diverse data will be combined. The test purpose and development plan will determine this. As discussed in 7.1.2, the IRT and panel data can triangulate each other in a number of ways, but only if the validation procedure permits this. For this reason, it is crucial that the pilot test cycle for new tasks and the main test cycle for existing tasks/tests are cyclical, so that IRT data collected on tasks which have been reviewed by the panel are then subject to review at the next panel event, leading to further recommendations for development which are implemented and subjected to further data collection, and so on.

8.2.9 Face validity

All qualitative feedback on task/test content from whatever source will inevitably include an element of face validity, and this is implicit in the communicative criteria discussed in 8.2.2.-8.2.7 above. Even expert panelists who have been instructed to look for specific criteria will be influenced by their reaction to the look and feel of the test as a whole. It may therefore be wise to elicit these views explicitly, by addressing a question to all stakeholders about their overall view of the test and its contents. These would be fed back into the next review cycle.

The reactions of test candidates are particularly important, as these may influence how hard they try; the extent to which they actually engage in the test is itself a measure of its validity. A short questionnaire might be prepared for candidates who have just taken the test or, in certain circumstances, to be administered a one to two months after they have taken the test, so that they can compare it with the TLU.

8.2.10 Empirical validity

Both IRT and expert panel can provide empirical data about such characteristics as task difficulty, item fit, skills tested, and rating scales. Some panel data, such as scrutiny of proposed tasks on paper, are 'test free' and can be generated even before tasks are trialed. This activity is particularly important for the pilot test cycle. Subsequent panel reviews include consideration of IRT and other task performance data as well as paper-based scrutiny.

IRT analysis can be used for adaptive test data where conventional test statistics are inappropriate, and will therefore make a major contribution to establishing the test's empirical validity. Because not all candidates take all tasks, however, a larger number of tests events needs to be administered in order to ensure a satisfactory sample size for IRT analysis, which in the case of a 'high stakes' test may imply the empirical analysis of a substantial pilot phase (at least several hundred) before the main test cycle can begin.

Subsequently, multi-faceted IRT can be used to identify the effect of different raters, rating scales, and language skills, but an even larger number of cases is required for this (of the order of several thousand).

Panel and IRT empirical data can be combined and used in a number of ways:

- a) to provide the 'hard data' for the individual task records in the task bank
- b) to focus on task reliability, by comparing item fit data from IRT with inter-rater consistency judgements from panel and rater training exercises
- c) to feed back into the algorithm design
- d) to exploit efficiency of the adaptive test to optimise use of resources, that is, to minimise the number of tasks and the time taken to achieve valid and reliable results.

8.2.11 Reliability

This could be considered as a specific form of empirical validity (Underhill 1987) but is sufficiently important to be treated separately here. Indeed it can be argued that of all the forms of validity, reliability is crucial in its own right to the overall validity of the test. Rating judgements need to be analysed for intra-rater and inter-rater reliability (consistency of judgements within and between raters) and a rater management system established to monitor this as a continuing process. The introduction of new tasks or new test forms will require checking for rater consistency among other factors. For each task separately, an estimate of task reliability would aim to identify the spread or distribution of scores for candidates with similar overall scores, and thus the extent to which assessors found it easy or difficult to be consistent in their scoring of that task.

8.2.12 Consequential and external validity

There is a very wide range of possible sources of consequential and external validity data, so much so that it may be difficult to assimilate or accommodate within the model of a test developed for a specific purpose within limited resources. It may be difficult to collect desirable data in certain circumstances, such as military or commercially sensitive contexts, or in cultures where the notion of test-takers as stakeholders whose views are important is novel. There may be little sympathy among stakeholders for the large-scale collection of consequential impact data. However, through the stakeholder feedback exercise, it should be possible to establish the impact on the most important groups concerned, the individual candidates, teachers, employers and co-workers.

Even where data can be gathered, it is likely that some of the data will be contradictory, where for example different stakeholders espouse differing and deeply-held views on learning, teaching and assessment; or the data may not be meaningful, where a new test is being developed precisely because there exists no satisfactory local alternative and perhaps little awareness of the need or scope for the new test. The issue of control of the use of test data was discussed in 7.1.4 and even where the administration of the test is carefully monitored, test developers can only advise on the interpretations put on test results by end-users.

Candidate scores on established criterion tests provide a valuable source of external concurrent and predictive validity, subject to the caution that correlations between scores only reflect a mathematical relationship between them, and many variables confound a simple validity interpretation.

8.3 The validation model

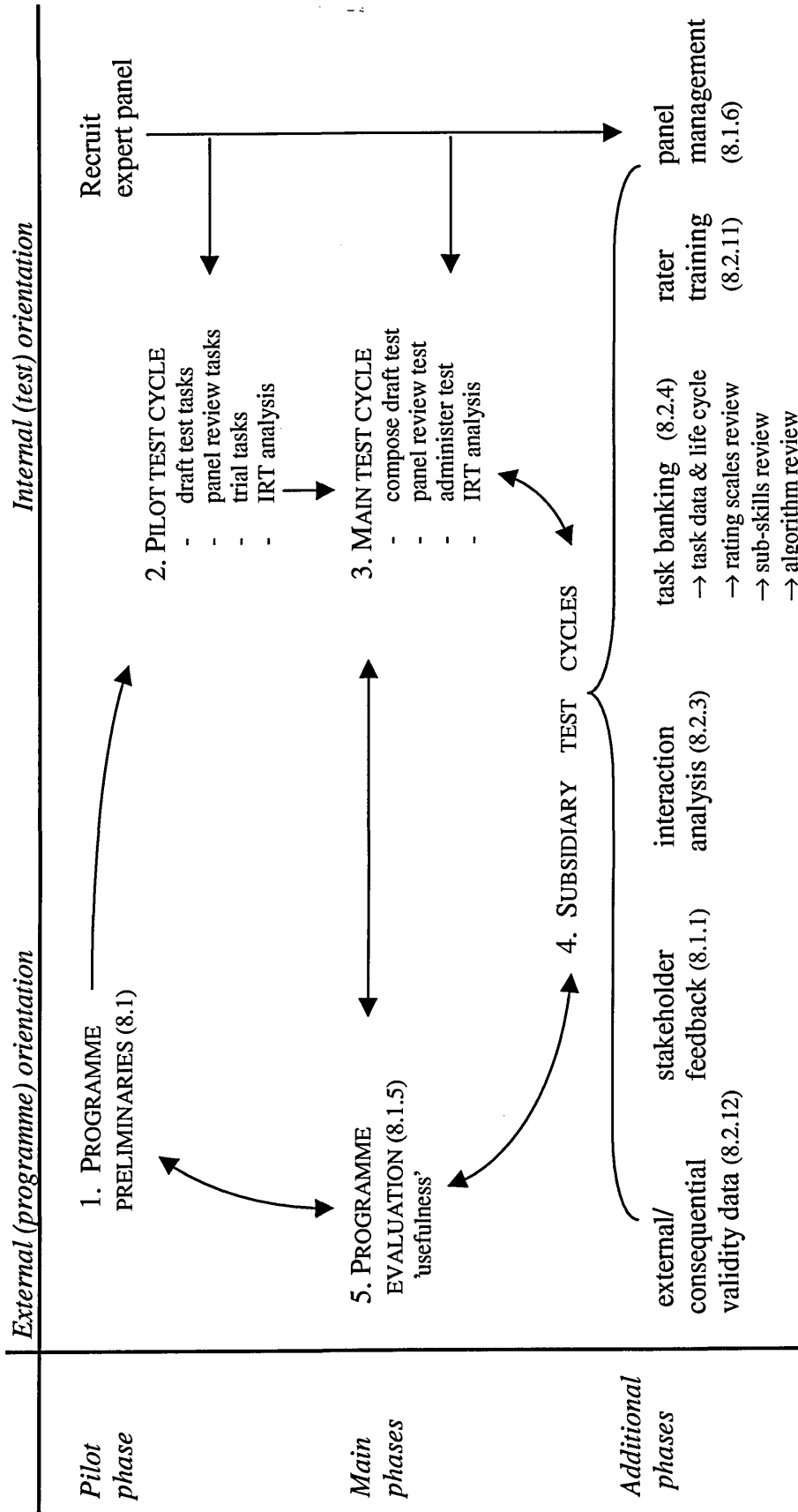
In this section, the model for the continuing validation of an adaptive language test is presented. It takes the characteristics described in the previous sections and combines them into a single overall model with a number of distinct components.

The model is founded on the axiom that validation must be a recurrent process rather than a single procedure, and therefore the model and each of its components is cyclical in nature.

As validation proceeds, and new evidence about the meaning of a test's scores becomes available, revisions may be needed in the test, in the conceptual framework that shapes it, and even in the construct underlying the test (APA, 1999: 9)

Figure 8 shows a diagram of the overall model, and this is followed by a summary description of its components and axes. The next section 8.4 gives further detail of the major sub-components of the model.

Figure 8 Diagrammatic representation of model for continuing test validation



Summary description of the model in Figure 8: axes and components

The model combines the desired characteristics identified above as a number of components on two axes. Both the model itself and its major components are cyclical; they are processes that need to recur continually as long as the test is in operation. The separation of validation activities into different components allows their cycles to operate at different rhythms, according to the demands of the test and the context. The two axes of the model present different phases of validation on the vertical axis and internal/external orientation on the horizontal axis.

On the *vertical axis*, the pilot phase includes the programme preliminaries described in 8.1 above, the recruitment of the expert validation panel independent of the test developers and the pilot test cycle for developing new tasks.

The main phase consist of the main test cycle, where the full operational test is delivered and the programme evaluation is carried out, which is an overarching evaluation that draws on the validation activities but remains external to the validation process. The additional phase comprises a total of six subsidiary test cycles containing further essential validation activities that feed off the main test cycle data and in turn inform main cycle validation decisions.

Although the pilot phase necessarily precedes the other two, there is a continuing interplay between the three phases, and the decisions made in the programme preliminaries will require regular review, like all the other components of the model.

The *horizontal axis* distinguishes the validation activities on the bases of external vs. internal test orientation. For theoretical and practical reasons, the validation of the test must be firmly embedded within the local context, and the model must reflect this. Therefore the programme preliminaries encompass the starting point of the test development as an external origin, rooted in an identified need in the local context, and similarly the programme evaluation is the external perspective that anchors the test programme to its broader context, whether that is commercial, educational or political or any combination of these.

Omitting the external orientation from the model would be unrealistic, in practical terms, but would also fail the criterion of having a real sense of communicative purpose. Only a test that was designed purely for the purposes of academic research might not have such an external focus, and in such a case the lack of communicative purpose could present a real difficulty.

The pilot and main test cycles draw on the external orientation of programme preliminaries and programme evaluation but focus on the internal workings of the test validation activities. The subsidiary test cycles have a mixed orientation; some, such as the task banking, rater training and panel management are internal activities, while others such as the stakeholder feedback and above all the consequential validity perspective are externally focused.

8.4 Further description of model components

8.4.1 Component 1: programme preliminaries

These activities are essential precursors to the test validation and are discussed individually in 8.1.1 - 8.1.6 above. The definition of test purpose, determination of target language use, initial stakeholder analysis, formulation of development plan are starting points deriving from external sources that inform the subsequent internal test cycles. The design of the algorithm (8.1.2) and the establishment of the expert panel (8.1.6) are internal activities, although consideration may need to be given to aspects of external face validity (e.g. individual panelists' status in the eyes of stakeholders).

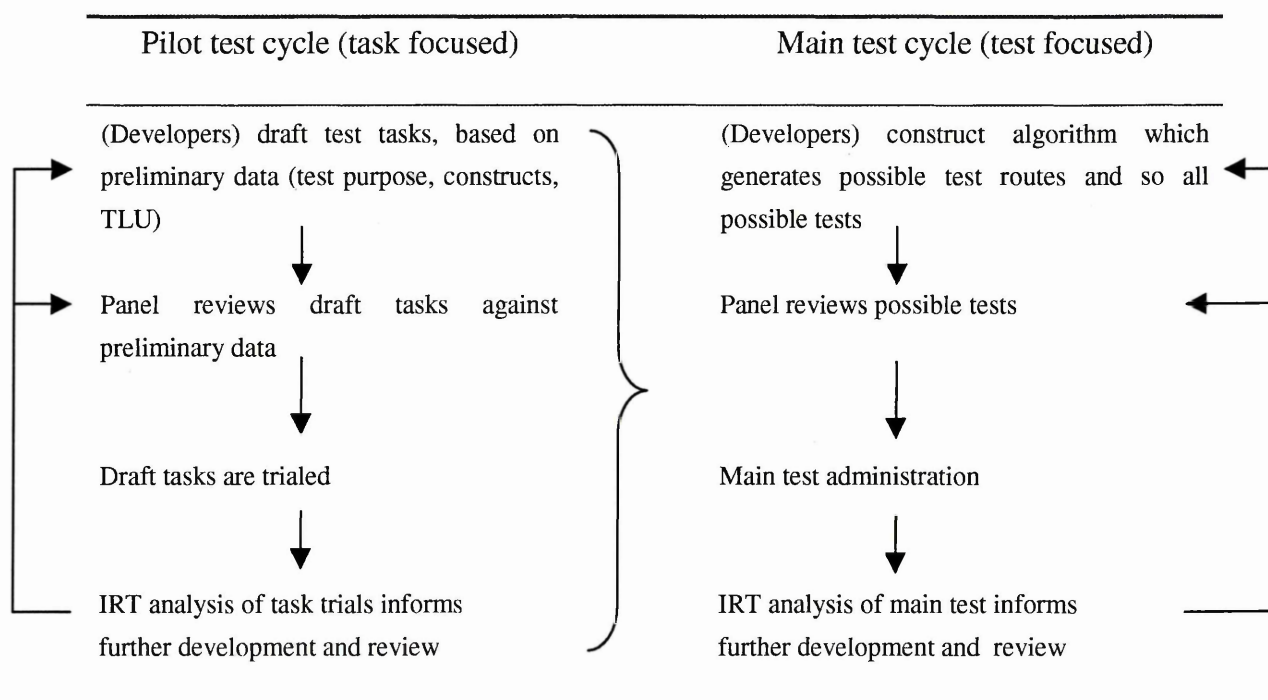
Setting up the programme evaluation is included here because it needs to be established at the outset, but it is discussed in more detail below. Ideally, it would link with the development plan to measure performance against specified objectives, which might include specific targets in areas such as numbers of candidates, income generation and stakeholder satisfaction. Test validity should also be among the objectives, although the formulation of quantifiable targets would be difficult in the current broad paradigm of test validation.

8.4.2 Components 2 and 3 : pilot test cycle and main test cycle

The pilot and main test cycles share a common four-stage framework, with a task-focus at the pilot stage being paralleled by a test-focus in the main test cycle. This is illustrated in Figure 9.

Figure 9

Common framework for pilot and main test cycles



This contrast of task- and test-focus in Figure 9 is idealised. The panel will also be concerned with the overall look and feel of the test at the pilot stage, and will necessarily review individual tasks in the main test cycle.

The panel's primary purpose is to review tasks and tests, and to take data from its own reviews and match them with data from IRT analysis and other subsidiary test cycle sources to evaluate the tasks and tests and make recommendations for improvements. Data from different sources may reinforce each other; they may directly contradict each other, in which case further research and collection of data is indicated; or most likely they will provide partial triangulation but with unique information from the different sources that needs to be pieced together like a jigsaw. For example, IRT might indicate that a task was slightly misfitting, and panel scrutiny of the task might suggest some

areas of concern about topicality or authenticity, and a working hypothesis that minor changes to the task addressing the concerns would reduce the extent of misfit could be formulated and tested out.

The exchange of data is in both directions. The data generated by the panel and IRT analyses provide information for subsidiary cycles also, primarily the panel management and task banking cycles. Other data generated in the task trialling and test administration activities may include:

- task and test scoring data, which feeds into rater training and task banking / algorithm review cycles
- audio or video recordings and participant introspection feeding into the interaction analysis cycle
- participant feedback, e.g. from survey responses, feeding into the stakeholder feedback cycle

8.4.3 Component 4: Subsidiary test cycles

Six subsidiary test cycles are located on the model as additional phases separately from the main test cycle. This is partly for clarity but also enables them to be seen as distinct activities which may follow their own time scale and which may be partly devolved. For example, the interaction analysis may require skills outside the scope of the development team and this may be sub-contracted to an external specialist working on a much longer time scale than the routine task banking cycle.

Panel management

The expert panel recruited as one of the preliminary activities has to be managed according to the agreed statement of policies and procedures. The continuity, maturation and mortality, in the sampling sense, of panelists all need to be considered in evaluating the consistency of panelists' judgements. The panel are not directly involved in test events in the way that raters are, and the cycle of panel activity is dependent on the presentation of new proposals by test developers and the analysis of test data by IRT.

As suggested in 8.1.6 above, routine operation of the panel's activities may be conducted by electronic means, which would greatly ease the process of recruitment of suitably-qualified panelists. Speed and thoroughness of response when invited to comment on tasks or tests would become key criteria for measuring panelists' performance.

Rater training

Raters are required for trialling draft tasks in the pilot test cycle as well as in the main cycle, and these test events provide the primary data for rater training as they do for task banking. Routine rater moderation would involve longitudinal scrutiny of each rater's judgements over a period of time to assess intra-rater reliability and cross-sectional comparisons of the different raters active at one time to establish inter-rater reliability. In addition to the analysis of rating data from routine test events, the rater training cycle should include special moderation events using recorded test data to focus on specific tasks or test aspects.

Like the panelists, the management of the rater training cycle needs to consider the recruitment and induction period as well as maturation and mortality in order to provide consistency and continuity of rating judgements. However, rater activity is directly linked to test administrations, and rapid growth in the number of test administrations would only be possible if there was slack in the rater capacity or a minimal delay in generating extra capacity.

Separate training and moderation for interlocutors is desirable, whether or not the interlocutor role is in fact combined with the rater role.

Task banking

Task bank data are generated by the pilot task trialling and by each test administration. Whereas data on task performance can be added to the task bank on an incremental basis, IRT analysis is routinely carried out as a 'batch' operation, at certain intervals of time or after specified numbers of new test events. Ideally, the computerisation of the test programme would save the data to the task bank automatically, so that any search of task bank records would always be based on the most up-to-date information. Manual data entry, whether individually test by test or in batch form, is tedious and prone to human error. Amongst other data that could be generated for review would be task difficulty, goodness of fit to the IRT model and task reliability.

Additional data on each task is provided by other validation activities such as panel review, interaction analysis and stakeholder feedback, and taken together it provides a rich source of information on each task. The developers and/or panelists would review

each task performance on a routine basis, after a certain period or a certain number of administrations had generated a pre-determined amount of data for IRT, or on an exceptional basis, where a potential problem had been highlighted by feedback from another source. In either case, tasks can be carried over unchanged, revised or abandoned.

There are three possible additional activities of the task bank cycle that provide further analysis of the data. These focus on the skills that are being tested, the scales being used and the algorithm underlying the test.

a) Skills facet Where there is a single construct being tested, there will be no need to identify different skills from the test data and IRT fit data will indicate the extent to which each task measures the overall construct under measurement. Where a test purpose sets out to tap into different skills, then IRT analysis can be used to test hypotheses about the extent to which different tasks measure different skills or combinations of skills. Subsequent main test cycles can then inform this construct modelling and refine the skills analyses of each task. This information in turn may be applied to fine-tune the selection of tasks by the algorithm and the weighting and allocation of scores across skills, for example where the test result is given as a profile rather than a single score.

b) Scales facet Where the assessment of performance on a task relies on rating against scales, which is typically the case for communicative tests, there needs to be a separate focus on how well these scales are themselves performing. IRT data can again be used to assess difficulty and extent of fit; in this case, it is each 'step' - the

level or rung on the rating scale - that is being analysed. Subsequent review can then revise the wording of descriptors or whole scales; a good example of this process is given in North and Schneider (1998).

- c) **Algorithm** The task bank data should also affect the task sequencing decisions made by the algorithm. If the algorithm is driven by random selection of appropriate tasks, then adjustments to the task bank data should immediately be reflected in algorithm decisions. For example, if a task is deemed to require a higher level of proficiency than originally envisaged, then that task should henceforward only be selected when previous performance dictates that a more difficult task is required. Scrutiny of test records will indicate how frequently each task is being called upon, and may show for example that there is a profusion of tasks available at one level of difficulty but a scarcity at another. As an alternative to creating new tasks, short-term action to remedy this might be to revise existing tasks to make them harder or easier.

If the algorithm has been developed by hand, however, then periodic adjustments to it will be needed to reflect changes in the availability and characteristics of the tasks in the bank. As the main and subsidiary test cycles generate new data, records on each task will be updated and task characteristics such as proportions of skills tested and difficulty estimates will be revised.

Interaction analysis

This subsidiary test cycle will take its raw data directly from the administration of test events in the main test cycle. As noted above, these data would typically take the form

of audio or video recordings of test events, which might then be transcribed and analysed to look for patterns among the use of different conversational strategies. Participant introspection about test events, involving both interlocutors and candidates, might also provide valuable data, and could at the same time be combined with the elicitation of stakeholder feedback on individual tasks.

A detailed transcription and analysis of even a single test event is a substantial undertaking and could only be undertaken on a very small sample of all tests. A more cursory scrutiny of the presence or absence of particular strategies, such as reformulation, could be undertaken on a larger scale and could also provide data for rater training and moderation.

Stakeholder feedback

It can be seen from the detailed diagram of the test validation model that the last two subsidiary cycles, stakeholder analysis and external/consequential validity, lie towards the 'external' end of the internal/external continuum.

Feedback from stakeholders may range from short-term comments on specific tests and tasks to which they have recently been exposed at one end to longer-term feedback on local perceptions of the test more generally. This is illustrated in Figure 10. Both types of feedback are useful, but the first type more so for internal test validation purposes, the latter for external programme evaluation.

The stakeholders whose opinions are likely to be most valuable for the purposes of validation are test candidates; test administrators; teachers whose students take the test; and end users such as employers, co-workers, or academic tutors.

Figure 10 Short- and long-term stakeholder feedback

Short-term specific	Long-term general
Short-term, exit-poll 'post-test' feedback on recently-administered specific tests and tasks	Long-term, more general feedback on how the test is perceived among the target and local communities
From test candidates and staff	From a broad range of stakeholders
Validation data feeds into task bank, rater training and panel review activities	Evaluation data feeds into external/consequential validity and programme evaluation

External/ consequential validity data

Although this area is currently regarded as essential for establishing test validity, in terms of the type of data collected it merges with the broader programme evaluation activity. The difference may therefore lie in the external/internal use to which the data is put rather than in a clear distinction between the activities involved. If the term 'stakeholder' is interpreted widely to include anyone affected in any way by the test, then there is a substantial overlap between external validity data and stakeholder feedback.

In addition to the feedback from immediate stakeholders, there may be more general wider consequences for the local training or education system and for individuals within it. This is now typically labeled 'impact' (Bachman and Palmer, 1996: 29) and its most obvious exponents are the washback effect of examinations on classroom methodology and teaching materials (McNamara, 1996: 23).

One concern that may have ethical implications is the use of a test for 'deselection', as a quick and easy mechanism to filter out a large proportion of applicants prior to other selection procedures being involved. Although there may be a genuine requirement for English language proficiency, if applicants are rejected solely on this basis before any assessment is made of their other skills, the test will come to be seen as 'unfair' with a negative impact on its consequential validity.

For validation purposes, external data feed into panel review of the test as a whole, and also into periodic review of the test purpose (8.1.1), target language use (8.1.3) and main constructs (8.1.4) itemised under programme preliminaries.

8.4.4 Component 5: programme evaluation

The facility for the overall evaluation of the model must be built into the development plan at the outset and the collection of data for evaluation must be a continuing activity. At the same time the framework for evaluation must be independent of the validation model itself in order to allow objective judgements about its performance to be made.

The evaluation framework suggested here is the basic model proposed by Alderson and Beretta (1992). It asks these questions about an evaluation:

- 1) Purpose: why is this evaluation required?
- 2) Audience: who for?
- 3) The evaluator: who?
- 4) Content: what?
- 5) Method: how?
- 6) Timing: when to evaluate?
- 7) Report: what and when?

These questions would be addressed in the present model at the programme preliminaries stage.

- 1) The purpose of the test programme will already have been stated (8.1.1); the central purpose of the evaluation is to determine to what extent the test is achieving its purpose. The local context may add additional purposes to the evaluation.
- 2) The audience will be defined by the stakeholder analysis (8.1.1). There may be different audiences for different types of reports.
- 3) The choice of evaluator depends on the circumstances. Some stakeholders may feel more comfortable with someone already associated with the local context, while others will prefer the perspective of a completely independent outsider.
- 4) What is to be evaluated will embrace all the objectives set out in the development plan (8.1.2). It is likely that test validity, in its widest sense, will be one of the objectives, and thus the validation process set out here will be central to the overarching programme evaluation.

- 5) How it is to be evaluated will depend largely on the value the stakeholders place on the evaluation, and thus on the resources that can be allocated to it. A small-scale evaluation can be based on a periodic scrutiny of key performance indicators, and may even be performed at a distance; this could include formative process-oriented as well as empirical data. A large-scale naturalistic evaluation would require the evaluator to be based in the local context and interacting with different stakeholders on a daily basis.
- 6) The timing also depends on the stakeholders and the development plan. If the programme evaluation is to parallel the test validation process presented here, then it too will recur on a cyclical basis. The timing of formal evaluations is often tied to funding reviews, and thus to the financial aspects of the development plan.
- 7) What to report and when: there may be different levels of reporting on different time scales and to different audiences, for example, internal and external to the test programme.

Further guidance on setting up the evaluation can be found in the literature cited in 8.1.5. In particular, Lynch presents what he calls a context-adaptive model (CAM) which is designed to be a "flexible, adaptable heuristic... that will constantly reshape and redefine itself, depending on the context of the program and the evaluation" (Lynch, 1996: 3). The 'context inventory' part of his model provides a checklist of potentially relevant dimensions on which to describe the programme, including the availability of a comparison programme in a similar setting, rather like a control group in a formal experiment.

Much of the data for the programme evaluation will already have been generated by the validation activities presented in the model here. It may even be that the programme

evaluation and test validation activities could share certain resources, notably research expertise. What is crucial, however, is that the externally-oriented evaluation remains sufficiently independent of the internally-oriented validation process to be able to determine, objectively, whether it is successful in its own terms and in terms of its contribution to the achievement of the objectives of the overall programme of which it is part.

8.5 Summary and limitations of the model

It will always be difficult to define with any precision a procedure that is intended both to be context-sensitive and to have a wider range of application. The model presented here is necessarily a compromise that draws on a critical evaluation of validation literature and the communicative methodology in earlier chapters, and at the same time, it has been elaborated with real-world constraints in mind. As an exercise in validation of the model, it is applied to the Five Star test in the next chapter.

The attempt to combine in a single cyclical model validity constructs from psychological testing with language teaching methodology and context sensitivity to local circumstances creates a number of barriers which act as limitations on the explicitness and applicability of the model.

Firstly, there are long-standing academic debates within the theoretical fields from which the model is derived which remain live, and these are carried over unresolved into the model. The extent to which the consequences of testing should be seen as part

and parcel of test validity is one example; the absence of an agreed rationale for combining language systems into a test is another.

Secondly, the practical application of communicative language testing methodology lags behind the theoretical discussion in the academic literature. Tests which claim to be truly communicative are not widely used, in part because performance tests require more resources than pen-and-paper or keyboard tests. Taken at face value, the suggestion that there cannot be an 'all-purpose' global communicative test because of the requirement for "the identification of test purpose and matching of tests and tasks with target language use" (Weir, 1990, quoted in section 2.10) means that truly communicative tests will never have the substantial resources for carrying out and reporting detailed validation studies. The effect is that there is still insufficient literature reporting on practical applications of communicative tests in different contexts to draw on in a model such as this, and none that claim to combine face-to-face interaction with adaptive computer delivery.

Related to the resource issue is that this validation model is essentially deterministic; it requires considerable forward planning to anticipate the direction and rate of growth of the test. For most new tests as for many other projects, the development plan is emergent - it is not possible to map out exactly what will happen in advance. The intention to commercialise and expand the use of a test outside its original target market is often not conceived until the pilot version has shown itself to be successful in meeting a local demand, as was the case with Five Star test, but it is in precisely these circumstances that the model for continuing validation, building on data collected from the outset, is most needed. A small-scale test developer seeking authorisation from

stakeholders to proceed with a small scale test pilot will be very conscious of cost constraints and will not want to burden their original business and development plan with apparently over-elaborate procedures for data collection and analysis.

The model does allow for some flexibility by setting up much of the activity in the subsidiary cycles, where the frequency of recurrence can be determined by the level of activity and other local constraints, more or less independently of the main test cycle. In some cases, cycles such as interaction analysis or stakeholder feedback can be set up on a modest scale to start with and upgraded in level of activity, precision and detail when scale of the test permits it and the resources justify it. However, all of the cycles require a base line to be established as rapidly as possible against which new data can subsequently be compared, and a failure to collect data systematically from the outset may only delay the point at which true longitudinal comparisons can be made. Other test cycles such as task banking require the initial establishment of a potentially complex data collection system requiring specialist expertise in the form of database programming and management skills.

Establishing the expert panel also requires a significant initial outlay in resources and a conviction that the greater credibility and validity of a truly external body can be justified. The panel provides a rich source of validity evidence that blurs some of the contrasts considered in the previous chapter; it can generate both qualitative and quantitative information, it can operate on both an evidential and consequential levels, and as an external panel it bring an external perspective to the consideration of internal test issues.

The value of this external expertise could be replicated on a smaller scale initially by the use of one or two expert consultants only, but their comments and recommendations cannot have the same authority as decisions made by a larger panel through a sequence of anonymous consensus. A smaller panel will reflect the views of each individual more prominently and underline the need for careful selection of external experts. Their impartiality is crucial to the success of the model, yet there will always be a tension between the view that their complete independence is essential and the opposing view that any degree of knowledge of the local context and culture will be valuable in helping understand the circumstance under which the test has to operate.

In either case stakeholders will quite probably want to influence the choice of these experts to ensure that their views are represented, and there is a risk that the panel becomes a group of partisan nominees for major sectoral interests rather than a truly independent group of experts. However the panel's views are collected, the process of formulating the questionnaire or interview instruments to avoid sources of bias wherever possible requires considerable experience in qualitative research methods. Where test developers do not themselves have such skills it may again be hard to justify the time and resources needed to establish the panel activities using sound instruments and analytic procedures at the very beginning of the validation process.

Once established, panel review of tasks can begin even before initial piloting with actual test events, but any kind of analysis of test data obviously requires a certain number of test events to have taken place. The model assumes that enough data will be available on which to base empirical validation, but stakeholders may require prior evidence of validity precisely in order to authorise the use of the test on a scale needed to collect

those data. This is true of any type of *post hoc* analysis of data, but for IRT analysis in particular there needs quite a large number of cases, of the order of several hundred, before the statistical model can begin to make useful estimates of the statistics described in section 6.3. If the test is highly adaptive, partial datasets are created by not all candidates taking all items, and this may further increase the sample size required for analysis.

- 9.0 Introduction
- 9.1 Preliminary activities and sources of data
- 9.2 Validation activities and sources of data
- 9.3 Application of the model components to the Five Star test
 - 9.3.1 Component 1: programme preliminaries
 - 9.3.2 Components 2 and 3: pilot test cycle and main test cycle
 - 9.3.3 Component 4: subsidiary test cycles
 - 9.3.4 Component 5: programme evaluation
- 9.4 Summary and limitations of the validation exercise

9.0 Introduction

This chapter takes the theoretical model for continuous test validation derived in the previous chapter and applies it to the Five Star test. This reflects the first aim of the research stated in chapter one, 'to validate the 'Five Star' computer based test of language proficiency within its immediate social and cultural context'

The structure of the chapter parallels that of chapter eight. In sections 9.1 and 9.2, the activities carried out and the sources of data generated are compared against the test characteristics identified in sections 8.1 and 8.2. This evidence is then collated in section 9.3 to compare the components against the model in section 8.3.

Finally, the limitations of the validation exercise are considered in 9.4.

9.1 Preliminary activities and sources of data

The preliminary activities identified in section 8.1 were carried out for the Five Star in Saudi Arabia in the early 1990s before the current project was conceived, and are described in Pollard (1994) with more general background in Robinson (1996), described in chapter four above. They are summarised in Table 37 against the list of preliminary activities from chapter eight.

Table 37

Preliminary activities and sources of data for Five Star test

Test characteristics from chapter 8	Activities and sources of data available for Five Star test
8.1.1 Purpose of test and identification of interested parties	<p>9.1.1 The purpose of the test was identified as meeting 'the need for an English language proficiency test for placing people in jobs and vocational training' (Pollard, 1994: 36) through the medium of English in the wider context of SDT's human resource activities in Saudi Arabia.</p> <p>The initial stakeholder analysis was confined to SDT's parent company, British Aerospace, and its many job applicants, but the scope widened to include other commercial clients with similar needs when the potential application of the test was appreciated.</p>
8.1.2 Development objectives, including commercial and strategic considerations; design of the algorithm	<p>9.1.2 SDT's status as a commercial entity required normal business plans to be prepared and operated internally. These have not been available for this research, but Pollard (1998b) suggests that it was positive feedback to the early prototype that led to it being operationalised while still in the pilot form, and it was then used on a larger scale by British Aerospace Manpower Resources Department in the recruitment of Saudi Arabian civilians to BAe. Development objectives were in effect being set year by year on the basis of current feedback and performance. Some concerns about the speed of this operationalisation are expressed in Pollard (1998a; 1999).</p> <p>The algorithm that drives the selection of tasks for each Five Star test event was written manually, rather than based on a random selection of tasks meeting appropriate criteria for skill and difficulty. This allowed the test developer to create routes where a single topic can be developed over a series of tasks (see example in section 4.2) extending the coherent interaction over more than a single task.</p>
8.1.3 Target language use	<p>9.1.3 A process of 'Population profiling' is described in Pollard 1994, in which around 70 questionnaires and interviews with target participants were administered in a range of professional and academic/vocational areas and a small number of tape recordings made of the actual use of English in the workplace.</p> <p>While the range of workplaces sampled in this way was limited, the target audience forms a highly homogeneous group in the cultural context in which it is based: young adult male Saudi nationals entering the workforce and seeking employment, and there is an assumption of a high degree of similarity of TLU among them.</p> <p>Many of the task topics in the pilot test were initially trialed in the population profiling stage (see section 4.2 above) to establish 'their accessibility to the test population in terms of the language sample they elicited, and the naturalness with which they merged into the dominant [focal] task' (Pollard 1998a: 5)</p>
8.1.4 Central language construct(s) and test specification	<p>9.1.4 The same activities used to describe the target language use also contribute to defining in Pollard (1994) the central language constructs and test specifications within the communicative competence tradition, drawing on the Canale and Swain (1980) and Bachman (1990) models. The test developer singles out strategic competence 'defined as coping strategies extending over pragmatic and sociolinguistic competence' as a particular focus for the test (Pollard, 1994: 40) and describes task types to operationalise this construct.</p>

Table 37 (continued) Preliminary activities and sources of data for Five Star test

Test characteristics from chapter 8	Activities and sources of data available for Five Star test
8.1.5 Programme evaluation	<p>9.1.5 Like the development plan in 9.1.2, there was no coherent plan for evaluation at the outset, but positive anecdotal feedback within the company (Pollard, 1998b), and a perception that the test matched the target language use domains better than the traditional teaching and testing programmes described in 4.1, led to the expansion of the test project and to the allocation of funding for the critical review carried out at Sheffield Hallam University in 1996 (Pollard, 1997). This review (reported in Underhill, 1997) deliberately took a broader rather than a narrower view of the validation remit, and was instrumental in the company's decision to proceed to develop the test commercially beyond the pilot stage.</p>
8.1.6 Establish expert panel	<p>9.1.6 The expert panel was established specifically for the purpose of carrying out the critical review, before this research and the model for continuing validation were conceived. The panel procedures established and the data collected are described in chapter 5 above. Considerable efforts were made to avoid possible sources of bias in the research design of the panel's activities.</p> <p>The panel was completely independent of the test developers and was able to comment freely on all aspects of the test and its tasks. This feedback has since been used to inform the development of test and tasks. There was however no scope in the critical review consultancy for a recurrent cycle of panel activity.</p>

Overall, the preliminary activities in the model were carried out in one form or another, but records of evidence are not available in some cases either because of commercial confidentiality or because the model is being applied retrospectively after the initial development has taken place.

9.2 Validation activities and sources of data for Five Star test

The validation activities carried out are summarised in Table 38, again listed against the checklist of characteristics for the model in 8.2.

Table 38 Validation activities and sources of data for Five Star test

Test characteristics from chapter 8	Activities and sources of data available for validation of the Five Star test
8.2.1 The overall model of validity	9.2.1 No explicit model of validity was stated at the outset, but the features of the test described in 4.3 above and the types of task employed are largely testing communicative performance, suggesting that the current validation paradigm associated with the communicative approach is appropriate. Diverse types of evidence are available, primarily from two rich datasets, the expert panel and the IRT analysis.
8.2.2 direct testing	9.2.2 A distinguishing characteristic of the Five Star test is that although the tasks test different skills, alone and in combination as evidenced by the expert panel analysis shown in Table 15, all the tasks are delivered by the computer mediated through open-ended, face-to-face interaction with the interlocutor. In this sense, it fully qualifies as a direct test. In the pilot form, however, the interlocutor must also combine the roles of test administrator and assessor, which reduces the level of attention s/he is able to pay to maintaining the normality of the interaction with the candidate.
8.2.3 Interaction	9.2.3 Although interaction was specifically excluded from the list of skills that the panel allocated to each task on the grounds given in 5.1.1 above, stage 3 of the panel exercise described in chapter 5 was devoted to gathering information about interaction in the form of an analysis of the strategies used by candidates in a series of video taped tests. All 12 panelists were thus able to comment on the same 11 video tests, with the results discussed in chapter 6 and tabulated in full in Appendices XIV - XVII. However, the data gained were of limited value - see 9.4 below.
8.2.4 Task-based structure	9.2.4 The task-based structure of the test was adopted from the beginning and transitions between tasks are emphasized by the presentation of new data on the computer screen. However, no comprehensive task banking was carried out at the pilot stage, and assessments of individual task characteristics, such as task difficulty and skills tested, had to be made on an intuitive basis by the test developers, as the panel was not set up until subsequently and the IRT analysis not carried out until much later on. The panel and IRT data subsequently provide objective evidence of task characteristics which has been incorporated in further developments.
8.2.5 Individualisation	9.2.5 The first source of individualisation identified in 8.2.5 stems from the algorithm underlying the selection of tasks. For the Five Star pilot test, the algorithm was written manually, in other words, the possible routes through the test were pre-defined by conscious decision rather than programmed to select at random tasks that meet certain characteristics. However, even with a relatively small number of branching routes in the algorithm, the number of possible different tests expands exponentially, and a modest bank of 73 tasks makes possible many thousands of unique test events. The second source of individualisation is the task itself. Of the Five Star tasks listed in Table 4, about 20% directly require the candidate to talk about themselves, their personal history or their opinions for example, 'What were the English classes like at school? (task 4-7School/study 2); 'discuss the pros and cons of large and small family sizes' (task 36-58 Speculation 1)

Table 38 (continued) Validation activities and sources of data for Five Star test

Test characteristics from chapter 8	Activities and sources of data available for validation of the Five Star test
8.2.5 Individualisation (continued)	<p>9.2.5 Individualisation (continued)</p> <p>The third source of individualisation is the interaction between candidate and interlocutor. As well as the explicitly individualised tasks mentioned above, the rubrics for other tasks encourage the interlocutor to probe personal attitudes, for example, 'Politely challenge the opinions expressed' (task 40-62 Speculation 3), 'Explore the candidate's ability to expand on his decisions' (task 69-113 Conservation). Even where the task rubric does not require this, the interlocutor may choose to follow up something the candidate has said.</p>
8.2.6 Topicality	9.2.6 - 9.2.9 topicality, authenticity, content and face validity are considered together. These areas were all addressed by the expert panel, although the characteristics were not specified individually. In phase 2 of the panel exercise, each panelist reviewed and commented on each existing task individually :
8.2.7 Authenticity	How can the content of this card [task] be improved? How can the presentation of the card be improved?
8.2.8 Content validity	
8.2.9 Face validity	<p>In phase 3, they were asked again for 'suggestions for improving existing tasks or topics', for 'suggestions for new tasks or topics' and 'general suggestions or recommendations for improving the test'. In total, around 280 individual comments were generated in this way.</p> <p>A further measure of content validity was the content comparisons against the IELTS test, Trinity College Grade Examinations in Spoken English and University of Cambridge main suite examinations.</p> <p>Feedback from test candidates and other stakeholders gave strong anecdotal evidence for face validity (Pollard, 1997, 1998a).</p> <p>The IRT data also provides a perspective on this broad area of test content through the measure of fit, which indicates the extent to which each task contributes to the overall score. However, it is not meaningful on its own, and only be interpreted usefully by scrutiny of the task in question and this is where the IRT and panel data can support each other.</p>
8.2.10 Empirical validity	<p>9.2.10 Empirical data were collected through the two major data sets, the expert panel and the IRT. The panel exercise yielded</p> <ol style="list-style-type: none"> estimates of task difficulty against an external scale (Table 19) an allocation of skills tested by each task (Tables 15 and 18) an estimate of the proficiency levels of 11 video tests (Table 26) an assessment of use of interaction strategies and of whether they contributed to or detracted from performance (Tables 22-25). <p>The IRT data set based on 460 pilot test events yielded</p> <ol style="list-style-type: none"> task difficulty ratings for the second and third exits on the three-point scale for each task, with estimates of standard error for each and measures of fit to the IRT model (Table 29) candidate proficiency ratings, again with estimates of standard error and measures of fit (Table 31) an item-ability map that reports the distribution of on the same logit scale as the task difficulty ratings (Figure 5) the same data can also generate individual learner maps to explore the response patterns of individual cases (examples in Figures 5 and 6)

Table 38 (continued) Validation activities and sources of data for Five Star test

Test characteristics from chapter 8	Activities and sources of data available for validation of the Five Star test
8.2.11 Reliability	<p>9.2.11 An indication of task reliability is given by the standard error estimate of the IRT scores for each item threshold (Table 29)</p> <p>Both panel and IRT datasets contain estimates of difficulty of each task, and the overlap between these 'item scores' from two quite different sources can be interpreted as a form of reliability estimate. Table 33 in section 6.3 show correlations of .89 and .92 between the estimates of difficulty. This is an acceptable correlation for reliability given the very different sources from which the figures come.</p> <p>Rater reliability The fact that most of the pilot tests used for the IRT data were administered by a single interviewer limits the possibility of inter-rater variation, but one rater reliability exercise was carried out using the limited data available. A group of 20 candidates were all tested by two interviewers on a test/re-test basis and an overall inter-rater reliability of 0.94 reported (Pollard, 1994). This is a highly creditable figure, but should be interpreted with caution, as the sample was small and the two raters involved were the two principal developers of the test who had been working closely together.</p> <p>The video rating exercise by the expert panel provides some evidence of the variation in ratings assigned to video tests by a group of raters, but the fact that they were using a single external rating scale makes it hard to compare their estimates with the actual test results reported as a profile.</p>
8.2.12 Consequential validity External validity	<p>9.2.12 The only source of external validity available for this research has been the content comparisons against established external tests. The test developers have collected some other information locally in the context of their development plans for a commercial version of the test. The history of the test, as described in Pollard (1997, 1998a, 1998b), clearly identifies stakeholder satisfaction with the early performance of the test as the principal factor in its expansion and further development.</p>

The informal origins of the Five Star test as a purely internal test instrument mean that there is little publicly available documentation about validation decisions made at the outset (9.2.1) or external and consequential validity (9.2.12) . However, the portability and availability of the test itself and video recordings of test events for panel scrutiny, the richness of the panel activities, and the test record dataset between them generated substantial and diverse sources of evidence for validation.

9.3 Application of the model components to the Five Star test

The model proposed in chapter eight is designed to be integrated into the development of a test from the outset. The two major constraints on its application here to the Five Star test are firstly that it is being applied *post hoc*, after the pilot phase has been completed, and secondly that it has been derived and is being applied externally, rather than providing the internal framework that underpins the test development. The raw test data available for analysis has therefore been cross-sectional not longitudinal, and recurrence of the iterative cycles of analysis, review and trialling has not been possible.

Having said that, the sample of data used in the cross-sectional analysis was substantial, comprising 460 test records, with outcomes for each task attempted and 11 video recordings of complete tests. The pilot version of the test was analysed task by task by each of the 12 independent panel experts, making in total several thousand judgements on the test as a whole as well as each task. While not constituting a full application of the model, this makes a sufficiently strong sample of the model activities to evaluate its viability.

The major components of the model in sections 8.3 and 8.4 are now considered in turn with respect to the Five Star test.

9.3.1 Component 1: programme preliminaries

As discussed individually in 9.1 above, the preliminary activities were carried out locally before this model was formally conceived. Although full documentation is not available externally, there is sufficient material (Pollard, 1994; Pollard and Underhill, 1996; Robinson 1996; Underhill 1997) to locate the test firmly within its local context. There is a clear statement of rationale for the test, identification of the target market and a commercial purpose and, through the 'population profiling', collection of primary data about target language use in that market. Crucially for this validation exercise, there was an elaboration of the theoretical basis for the test within communicative competence approach, with a focus on strategic competence that underpins the central role of interaction between candidate and interlocutor in every task. This makes possible for the purpose of this research the use of the broad professional consensus about the key features of the communicative approach to teaching and testing as criteria for validation of the test.

9.3.2 Components 2 and 3: pilot test cycle and main test cycle

Because of the history of the development of the Five Star test, the expert panel was not established until the pilot test had been constructed and trialling was well under way. The explicit purpose of the panel was to critically review the test and to make recommendations for its further development.

Many comments related to the memory load, e.g. 'A memory task - not very interesting' (comment on task 22-29 Fridge) or the information load, e.g. 'occasionally candidates seemed to be overwhelmed by the quantity of information to process' (general comment); or the knowledge of the world required for processing tasks, e.g. 'it tests world knowledge as well as English - lapse may be due to ignorance of world affairs rather than English' (comment on task 42-65 Regional affairs)

Comments on topicality included 'this is going to date, and needs replacing soon' (comment on task 33-55 Kuwait City); 'this one has a current affairs angle which, as it becomes historical, will increase the difficulty of the task' (comment on task 34-56 Nagorno Karabakh).

Examples of other panel comments:

'A more authentic follow-up task would help' (task 26-34 Traffic lights 2)
'How could basically the same task be contextualised?' (task 5-8 Basic reading)

More sample panel comments are given in Table 20.

These recommendations were then drawn on in the development of the full commercial version of the test, which has only recently been completed (spring 2000).

In the Five Star case, the expert panel and IRT analyses have therefore been 'one time' activities taking place between the pilot and main test cycles, rather than recurrent activities within each cycle as envisaged by the model. The individual sources of data were described in chapters 5 and 6; section 6.4 in particular explores the possibilities of triangulation of data from distinct rich data sources, taking comments from the panel

exercise for example to explain why some tasks show up in the IRT analysis as misfitting and to suggest some possible lines for further enquiry.

Overall, however, less than 10% of both the task and the candidate scores reported outfit means squares of above +2, the limit suggested by Stansfield and Kenyon (1996) for measuring misfit. This suggest that the Five Star test data does indeed fit the IRT model.

9.3.3 Component 4: subsidiary test cycles

The subsidiary test cycles also vary in the extent of data available for Five Star validation.

Panel management

The comprehensive expert panel exercise was described in chapter five. The exercise was carefully structured and precautions were taken in the planning stage to ensure the independence of panelists' decisions and to guard against possible sources of bias. These precautions included:

1. The use of the Delphi procedure to allow panelists to contribute anonymously to the group consensus without being unduly influenced by their peers
2. The panelists were divided at random into two groups who analysed the Five Star tasks in a different sequence, to counter any order effect
3. Similarly, the two groups of panelists analysed the video tests in a different sequence

4. The panel completed the proficiency level and skill allocation assessments of each task before any exposure to actual test events on video, to prevent their being influenced by the performance of particular candidates
5. The preliminary round of panel activity was designed to produce agreement on the skill definitions to be used and so promote consensus on skill allocations in subsequent rounds

Rater training

A single interlocutor/interviewer carried out all the tests in the pilot phase. However, anticipating the need for training of multiple raters at the commercial stage, the critical review included an appendix (Appendix V) proposing a framework for training and licensing Five Star assessors (Parker, 1997). Twelve interviewers have now been trained to use the new version of the test, and a substantial Assessor's Manual produced for this purpose (SDT, no date) to form the basis of a three-day assessor training course. This includes detailed practical advice on the conduct of the test as well as a background to the instrument and the fundamental principles of performance testing.

Task banking

The expert panel and IRT analyses yielded substantial quantities of data on the Five Star pilot tests, which have been available to the test developers for subsequent test revision and development. These data are contained in the appendices where they are ordered by source (type of data and analysis) rather than by task, but they could easily be collated to pull together all the information on each task. This would create the core of the task bank.

The Five Star test is adaptive, but the computer algorithm underlying the process of task selection was written manually, and therefore does not fully exploit one of the potential advantages of adaptive tests, which is the random selection of tasks of appropriate difficulty. Task selection is therefore predictable if you know the outcome of the preceding task, and the deletion, addition or adaptation of any tasks will in theory require a manual reworking of the algorithm to reflect this.

Interaction analysis

As noted in 9.2.3, data on candidates' use of interaction strategies was collected as part of the panel exercise. This was based on a count of instances observed in video tapes of tests, rather than the preferred method suggested in 8.2.3, the transcription and analysis for turn-taking, topic nomination, length of turns and other discourse features, which would have generated much more detailed data but also required much greater resources to analyse. Such work is however now being carried out on Five Star test data (Pollard, in progress). The preliminary data from the expert panel analysis reported in 6.1.3 above identified tokens of interaction as a significant and observable factor in candidate performance, with some relationship between strategy use and overall level of proficiency. A separate analysis of these data by the Five Star test developer has influenced task design in the revised version of the test (Pollard, 1997).

Stakeholder feedback and consequential validity

Without ongoing access to the local context, this project has not been able to gather data in this area at all. As described in 9.1 above, it was a positive anecdotal response from both candidates and other stakeholders to the early use of the test that led to its expansion and further development, but no systematic evidence is available.

9.3.4 Component 5: programme evaluation

Again, this component of the model is being proposed retrospectively with regard to the Five Star test, and although evidence of the 'usefulness' of the test, in the sense of Bachman and Palmer (1996) model, has been gathered locally, an evaluation of the overall programme is not available for this research. That the test developers were satisfied with the performance of the test at the pilot stage, and with the tenor of the critical review (Underhill, 1997), is clearly indicated by their decision to proceed with the development of the full-scale commercial version of the test; indeed, concern has been expressed that it may have been 'prematurely operationalised' (Pollard, 1998b, 1999). In doing so, they will have considered issues to do with test delivery which are essentially practical but have implications for consequential validity, such as the ethical / commercial consequences of moving into 'high stakes' testing (Pollard, 1997) and the risks of possible test misuse outside the tightly controlled context of the pilot test. Options for promoting test security include a licensing system that allows only trained operators to administer the test, a requirement for the return of test data for analysis, and a limit on the number of test events that can be delivered before such analysis is fully evaluated.

The overall conclusion is that despite some quite major gaps in the data available for the validation exercise, particularly from external sources, the pilot test was a substantially valid test of candidates' communicative competence in English in the specific cultural and linguistic context within which it was used. Limitations and qualifications to this conclusion are considered in the following section.

9.4 Summary and limitations of the validation exercise

This chapter has taken the validation model proposed in chapter eight and applied it to the Five Star test, both to establish whether it will work in practice for a particular test and to explore its applicability to new features of tests that are likely to become more common, alone or in combination, such as communicative methodology, an adaptive algorithm and a computer-based delivery system. The datasets generated in the course of this research have produced large amounts of evidence from diverse types that can be pieced together to provide substantial evidence of validity, and this is very much in line with the current validity paradigm. However, the exercise has also thrown up a number of constraints.

The major limitations on the use of the model to validate the Five Star test stem from the relationship between this research and the test development. The model is designed to be integrated into the planning and delivery of a test throughout its active life, with recurrent cycles of activity reflecting continuing change to the test and its context of use and continually refreshing and reinforcing its validity in that context. While the periodicity of the different cycles will be determined locally and may fluctuate according to the pattern of use of the test, it is the recurrence of panel and IRT analyses that generates the data to drive the main and subsidiary test cycles. In the case of the Five Star test, this was not possible, and the data has been drawn only from one period, during the pilot test cycle. It is not therefore possible to carry out the recurrent test cycles and to report on their operation in practice; however, a single complete cycle has in effect been carried out.

A second consequence has been the 'back to front' nature of the test/validation relationship, with the pilot test driving the critical review and subsequently the model developed in this research, rather than the other way round. The research project and the model were conceived after the pilot test was already in use, and the model was formulated after the initial panel exercise had been carried out.

A third consequence has been the lack of access to some of the documentation that is internal to the test developers, commercially and geographically remote from the locus of this research. As the Five Star test was intended as commercial activity, some of the development documentation has not been as widely available as it might be for a test developed in a purely academic environment. It is in the area of external orientation on the model that sources of evidence for the Five Star test are particularly lacking.

Other limitations spring from some of the specific validation activities called for in the full model.

One interlocutor carried out almost all the tests analysed in the pilot phase, and so no serious comparisons of scores between raters or rater training activities have been possible.

Allocation of tasks to different skills was one of the major foci of the panel exercise, but the number of test administrations submitted to the IRT analysis was insufficient to differentiate by skills. Considerably more data would be required for this, of the order of 5000 test events rather than 500. In effect, the IRT analysis treated the data as representing a single facet only and implying a single language proficiency construct,

whereas the panel exercise clearly indicates a substantial measure of agreement between panelists that most tasks were tapping at least four distinct skills, separately or in combination.

The accumulation of further IRT data on each task would eventually enable these language skills to be treated as separate facets in the analysis, so providing a further cross-check between the IRT and panel analyses. However, this would require at least a working solution to the vexed issue of the status of interaction; would it be treated as another skill, comparable to listening, speaking or reading in the skills facet? How can such a position be justified in the light of current views of interaction as something unique to each event and co-constructed by the participants in that event?

The type of interaction analysis used in this research only allowed fairly general conclusions to be drawn about the use of these interaction strategies. A crude system of scoring the strategies observed on an adaptive test where different candidates attempt different tasks does not generate comparable statistics, and a more comprehensive analysis of the actual transcription of the interaction might prove more fruitful.

In addition, the results of the interaction analysis showed a wide variation in panelists' scoring and suggests the need for care in standardising and moderating panelists' judgements in this respect.

The 12 members of the expert panel were completely independent of the test but mostly unfamiliar with cultural context of the test; three of the twelve panelists had lived and worked in the same or similar cultural context. Argument could be made both for and

against recruiting panelists with prior knowledge of the cultural context of the test; if that experience was at any distance, geographically, temporally or culturally from the immediate time, place and context of the test, then it might be seen as introducing an extra source of bias into that person's judgements. On the other hand, there is no doubt that panelists who were completely unfamiliar with the culture found greater difficulty in understanding the context and purpose of the test.

The qualitative external comparisons in the critical review, which contribute to the concurrent validity, were only carried out against UK-based global tests of spoken English. These are useful for generating checklists for possible content areas but do not allow strong claims for criterion validation, because they are aimed at such a completely different market. The problem here is that finding an established exemplar to use as criterion test will always be difficult when a new test is being developed precisely because no current test exists that does the same job satisfactorily.

- 10.1 Comparison against aims
- 10.2 Key issues raised
- 10.3 Appropriate contexts for application of the model
- 10.4 Recommendations for development

This chapter summarises the scope of this research against the five aims stated in chapter one and identifies the key theoretical and practical challenges arising from it. Sections 10.3 and 10.4 describe a range of suitable contexts in which to apply and develop the model.

10.1 Comparison against aims

The first aim was to design a model appropriate for an adaptive test of spoken language proficiency that allows for the validation to become a recurrent process as the test evolves rather than a single procedure. This is done explicitly in chapter eight, with a series of different components drawing on diverse sources of validation evidence and different cycles allowing for responsiveness to local timescales and circumstances.

The second aim was to identify the distinctive features of the communicative approach to language teaching and testing and to discuss their implications for the model for continuing validation. These features are elicited initially through the historical overview in chapter two, which emphasizes the intimate relationship between language teaching and testing and the evolutionary nature of their methodology. The salient characteristics of the communicative approach are then developed in more detail in chapters 3 and 7, as discussed in the literature on testing validation and as implemented in practice in current language tests.

The third aim was to try out the model to validate the 'Five Star' computer-based test of language proficiency within its immediate social and cultural context. This draws on the description of the Five Star test and its context in chapter four and the collection and analysis of two substantial datasets in chapters 5 and 6, leading into the detailed application of the model to the Five Star test in chapter nine.

The fourth aim was to discuss the implications of the first three aims to explore a procedure that can be applied elsewhere for the validation of language proficiency tests that share some or all of these key features. The structure of this dissertation has been designed to allow continuing reference throughout to contexts and constraints other than the Five Star test against which the individual components and cycles of the model is systematically compared. Sections 10.3 and 10.4 below consider a range of test environments in which the model may be appropriate.

The fifth aim was to contribute to and enrich, at a theoretical level, the academic debate on issues surrounding language test validation. The central conundrum for language

testing in the real world is how to apply in practice the theoretical requirements for test validation generated by developments in validation principles and language testing methodology. Testing is almost always highly constrained by resource limitations, especially human resources and time available, and full implementation of the communicative methodology in the current validation paradigm is resource-intensive and context-sensitive. In practice, therefore, most existing tests cannot claim to be fully communicative. This research has attempted to take forward the theoretical discussion by developing as a framework a flexible, overarching model for continuing test validation that is founded on theoretical principles yet is applicable in practical everyday contexts.

10.2 Key issues raised

This model has been elaborated to meet a series of challenges, both theoretical and practical, which require test validation to become a continuing process. There is a tension between this and the reality that a test must at some point be considered sufficiently valid to be used with confidence to make decisions with consequences in the real world:

Unlike the researcher, who can afford to investigate the issue over a period of time, test developers need evidence of the validity of their instruments as quickly as possible (Alderson et al., 1995: 175).

A summary of these challenges will help to focus the applicability of this model to other tests and testing environments in the next section.

The theoretical challenges comes from the current validity paradigm, that evidence should be sought from a wide variety of sources, both empirical and naturalistic, all of which can contribute valuable evidence. Some sources of information for validity will be available at the inception of a test, but others will not become available until the test programme is under way, and even then will be subject to continual updating as new data comes in. Rather than a single stage that a test has to go through to establish its value, validation is now seen as a continuously-developing activity paralleling the use of the test itself, and this is why the model is based on a series of cyclical processes.

Validity generalisation is the extent to which evidence of validity based on one situation can be generalised to a new situation without further study of that new context (APA, 1999). This may in part be justified where it can be shown that variability in test scores is due to statistical procedures, such as sampling procedures and reliability issues, but in principle even minor variations in the use of a test need to be underpinned by continuing collection and analysis of data. One such minor variation is the continued use of a communicative test over time, when the dynamic interplay between topicality and authenticity of tasks and the real world contexts of the participants has a continuous impact on test validity.

A related theoretical development has been the shift in emphasis from the search for evidence of test validity as a once-and-for-all status attached to the test instrument to the contribution of this rich information to the interpretation of specific scores. A test only has meaning when its scores are used for some purpose, and it is the use that the test scores are put to that is or is not valid. Therefore any change in the use made of test scores and the kinds of decisions based on them requires further evidence of validity.

This is where the consequential aspect of validity reaches far beyond the empirical test data, and it is why the model needs to have an external as well as an internal orientation.

In order to validate any such changes in the context of use or candidate profile, the test developer would need to be able to compare the test, task or candidate data generated with previous data to determine whether in fact there were any significant changes. This implies the existence of baseline data for comparison from the outset and this is why the model presented here is integrated into the test development programme from its inception.

As well as conscious decisions to apply a test to a new market, there may be gradual demographic or socio-economic shifts in the candidate profile which are imperceptible on a day-to-day basis and go unnoticed. The collection and analysis of data therefore need to be a routine and continuing activity rather than a special event called into play at a particular stage of test development.

A further set of challenges is posed by the integration of a spoken element and the communicative methodology which dominates current language teaching and testing. While there is no consensus about the axioms of the communicative approach, there is a central requirement for consideration of factors such as authenticity and topicality which are not constant over time, even when other variables such as candidate and context remain unchanged. Any test with aspirations to communicative status must engage validation as a dynamic process on these grounds alone.

The inclusion of a direct speaking component brings in new issues associated in the literature with performance testing. Among these is the vexed status of interaction, which as a construct spills over from a purely linguistic skill to a social and psychological factor and from a variable that can be measured uniquely in one candidate to a by-product of any verbal exchange between two or more people that is co-constructed by all the participants. The assessment of interaction as a strategic skill displayed over a series of verbal exchanges may also be problematic in a test that is rigidly task-based with an assumption of independence between the performance on consecutive tasks.

Adaptive testing does not depend on computer delivery nor do computer-based tests need to be adaptive; arguably it is direct tests of spoken performance with a live interlocutor that can be said to be the real pre-cursors of adaptivity in communicative language testing. Recent technological innovation has made possible the storage and delivery of language tasks in an electronic form following an adaptive algorithm of task sequencing. The relatively low cost of the hardware and software now allows tests to be continuously updated and duplicated at little additional cost compared to the economics of print-based publishing. The challenge here is to combine the technology and the methodology, to determine how best to exploit the potential of this delivery system with the key features of a communicative test. So far, few language tests appear to be responding to this challenge. Given the evolution of web-based language tests referred to in section 4.5, it seems likely that this will be another medium for exploring how the technology and methodology can combine. Can genuinely communicative language tests be delivered at a distance over the internet?

10.3 Appropriate contexts for application of the model

The model presented here has been developed for the validation of tests with a combination of some or all of these features in mind:

- adaptive
- computer-based
- communicative, including authentic and topical test activities and materials
- task-based
- with a component of direct speaking involving interaction between candidate and a live interlocutor
- tapping integrated skills in the course of test events
- aimed at a distinct target market with a target language use domain that can be described with some degree of precision

In its present form, the model does not prioritise among these features nor discriminate optional ones from those that are criterial. One can speculate that, in general, the more of these features are present, the more suitable the model will be. Existing tests which explicitly violate one or more of these features may not allow the operation of the model in its present form; either the model can be adapted, as envisaged in the following section, or the existing test can be adapted to fit better. In general, it is likely that this will bring the test more into line with current communicative practice.

For example, tests which are not task-based but consist of a single extended performance will not be immediately amenable to the kind of analysis suggested in this

model. Three examples are an extended oral interview for employment or academic entry, in which language proficiency might only be one aspect under investigation; an open-ended essay or composition; or the presentation of a project which does not easily break down into distinct sub-tasks. Doubts about scoring reliability and equal opportunity for diverse candidates surround such 'single task' tests, and a move towards a 'multi-task' format would ameliorate these concerns.

The model is in principle equally applicable to tests which are not adaptive. If they are still item- or task-based they will normally generate complete data-sets with all candidate taking all items, and so will be able to use conventional test statistics such as internal reliability coefficients and item facility and discrimination indices to analyse test results. With the benefit of this empirical evidence there would be less urgent need for IRT treatment, but it would nonetheless still be useful to carry out IRT as well as the traditional statistical analyses to provide independent item and candidate estimates for trialling the test in new contexts and with new populations.

Some examples of common test purposes where the model could be applied are the English language component of a recruitment and selection procedure for a large private or state enterprise with a significant requirement for English language skills among its workforce; an achievement test for a credit-bearing language course at a tertiary level institution; and a local or regional school system seeking to modernise its English language teaching and testing within the broader curriculum framework.

There are a number of aspects of the broader environment of testing programmes that bear on the suitability of this model.

1. The model would be particularly suitable for highly dynamic situations where the target market, the test contents and the use to which test results are put are likely to change over time. Commercial applications will seek to create and exploit new markets. National, regional or local educational testing systems have to be accountable, ultimately to their consumers; language testing never takes place in a vacuum, and there are social, political and demographic changes constantly taking place in every society. The growing role of the English language in particular as a medium for international communication is accelerating change in the weight given to English language assessment in all educational systems.

By implication, the model would be less applicable to tests in static contexts, such as a test developed solely for internal use within an institution that would never be used for any other purpose. Even so, any kind of evaluation that is carried out on the test programme is likely to produce some recommendations for improvement, and these will be easier to identify and implement if they result from a systematic and continuing validation process with baseline data for comparison. Even on a local scale, demand for language teaching and testing can change dramatically, and tests that were at the outset conceived as small-scale and limited in scope may quite rapidly be expected to serve a wider constituency, as the example of the Five Star test shows. Psychometrically naïve stakeholders who have been impressed with the feedback on test performance may have little sympathy with arguments that the test was never designed for wider use. A systematic validation process will provide the data at any time to allow that extension to take place quickly yet on the basis of solid baseline evidence.

2. The model is suitable for situations where there is local responsibility for test design and decision-making, or at least a strong local influence on design choices, even if there is a subsequent stage of seeking approval from stakeholders who are more remote. In commercial contexts, this is unlikely to be a problem, and responsiveness to feedback and concern for consequential validity should be seen as good business practice. Stakeholders need to be open to persuasion of the value of establishing a long-term process of continuing validation, ideally within the evaluation of any greater programme of which the test forms part. The model provides a sequenced framework for implementation which may actually help to create confidence among stakeholders that the development plan is well-founded.

The communicative requirement for context-sensitivity makes it impractical to carry out standardised communicative testing on a large, centralised scale. This has created a dilemma for external agencies such as ministries of education or central examining boards which have traditionally dictated the teaching syllabus nationally, and which now seek to introduce more communicative practices. Under pressure from teachers, parents and students a common response has been to devolve responsibility for teaching and testing to institutional, local or regional authorities, which would facilitate the application of this model. However, even in highly centralised systems, all the sources of data and test cycles remain valid in principle and test developers will still have opportunities for data collection and programme evaluation. In fact, it is the consequential data that can provide solid evidence of the need for significant change in the test system.

3. The model is suitable for tests being developed in all but the smallest language teaching or training programmes. The emphasis on test validation as part of a wider programme evaluation identifies it as an integral component in a larger system, with the possibility that allocation of resources to data collection and analysis can take account of parallel activities in the wider programme. For example, where a language test is needed in an industrial training context, much of the test specification and programme evaluation work will fit alongside similar activities that need to be carried out for the training programme as a whole. The cyclical nature of the model can be extended to form part of a broader institutional communication strategy, where the external and consequential validity data is derived from other departments within the organisation, such as job specification and feedback from line managers and co-workers on the performance of new recruits. Such prior work analysis and performance feedback could if systematised properly be linked directly into the test validation cycle.

Small scale testing programmes with limited resources will have difficulty putting all the components of the model in place at the outset. In such cases, there is a tension between the desire for a systematic model to be established at the outset and the need for a flexible process that can reflect and respond to the evolutionary nature of test projects. At the practical level, there may be a problem identifying and committing sufficient resources at any early stage to set in place a systematic validation process at the outset; many testing projects, like Five Star, start in a small way and gather momentum as positive initial feedback indicates a need is being met. It may only be then that the real potential of the test is revealed, and so only then

that the necessary resources can be justified. Although it is not ideal, it is possible to introduce the model at any stage and apply some components retrospectively.

4. Because of the feedback loop created by the inclusion of continuous cycles of operation and consequential validity data the model is suitable for tests that are used for medium or high stakes decisions, where the consequences of either random error or systematic bias in results may be severe. In the short term, the model is responsive through the built-in system of pilot, main and subsidiary test cycles. In the longer term, the selection of panelists and the extent of involvement of the expert panel can be used to respond to concerns about the representation of expertise in specific areas such as the end-user perspective, target language analysis or psychometric theory.

Tests involving 'low stakes' decisions may not be able to justify the application of the full model. A test programme may be quite large, in terms of number of test events administered, yet the consequences of individual error not be severe. An example of such a case would be the use of a proficiency test to place students in classes in large language teaching programme, where mis-placements can easily be rectified by transfer from one class to another. Nonetheless, such transfers involve administrative resources, reduplication of assessment and possibly loss of face for the individual concerned, and a test that was consistently inaccurate would rapidly lose the support of stakeholders, and even 'low stakes' tests still require validation. It may simply be that the different stages of the model are applied in more informal and less fully documented ways.

5. Because of the high level of accountability that is built into the model, it is suitable for situations where test operation and results may be subject to public scrutiny and criticism. The routine use of external panelists in particular provides a robust element of independent accountability that is more likely to satisfy formal quality assurance arrangements than the use of purely internal monitors. The collection of detailed evidence that is required to operate the model fully can be used to provide a permanent audit trail for the investigation of complaints or allegations of misconduct or unfair practice.

Tests which are used for commercially, politically or militarily sensitive decisions may not be able to use an external panel of fully independent experts. Local experts may not be considered sufficiently independent, with all kinds of possible connections with candidates and their sponsors, and 'foreign' experts may not be trusted by stakeholders to have sufficient understanding of or empathy with the local social and political context. Typically, in such contexts external experts may not be consulted at all, with test development staff acting as the moderating panel in a more or less formal capacity. Nonetheless, the significance given to the external panelists in the model can be used to justify some element of external involvement, for example, of staff in other branches of the organisation who have the necessary affiliation or security clearance.

Overall, however, the model is designed to be highly flexible, with multiple sources of data feeding in, and in principle 'the more different types of validity that can be established, the better, and the more evidence that can be established for any one type of validity, the better' (Alderson et al., 1995: 171). The different components can be

adapted to draw on different sources of data, and the main and subsidiary test cycles can operate on different rhythms, allowing the model to be used in a wide range of possible test contexts.

10.4 Recommendations for development

The model has been tried out in this research against the Five Star test, and now needs to be validated with other tests in other environments. Any reports on such trials will generate feedback on the design and operation of the model for that test which may be applicable in other contexts also.

Full application of the model will provide a comprehensive and detailed blueprint for test validation from first inception through the pilot stages to the full implementation of the test as long as it continues to operate. At the very least, the model offers test constructors and developers a checklist of validation components and activities.

Between these extremes is a range of contexts where the design of the model provides a flexibility of application through the diverse sources of validation evidence sought and the independent timescales of operation of the different cycles. The value of each part of the model needs to be established through a kind of cost/benefit analysis, asking about the cost in terms of resources and the benefit in terms of validity information for each component. The cost of setting up and operating an external panel, for example, needs to be justified in terms of the independence of judgement and perspective it brings.

A more significant version of the checklist approach would be its use as a tool for resource planning and negotiation with stakeholders, especially where multiple perspectives (Lynch, 1996) are encountered. By showing the return on early resource investment in the form of later validity evidence, the model can be presented as a *modus operandi* with which to negotiate the allocation of resources and responsibilities.

Feedback from application in other test environments will show how robust the model is to permutations of the key features listed in the previous section. It was hypothesized that in general, the more of these features are present, the more suitable the model will be; common sense will suggest where modifications are appropriate. If an oral test does not aim to assess integrated skills, for example, there may be no need for an explicit skills allocation activity for the panel; however, content/construct evidence will still need to be collected that the test is tapping all and only the target language skills.

Repeated application of the model in different contexts might generate a hierarchy of key features. A commitment to communicative methodology, for example, might be found to be more necessary to the functioning of the model than an explicitly adaptive format or the use of computer-based delivery.

A further avenue for developing the model would be to explore its linkage or integration into larger institutional information systems. It was suggested in the previous section that the model could be successfully operated within a larger commercial environment with an established system of setting objectives and collecting data to measure performance against those objectives. An analogous context would be an academic institution with a requirement for assessment strategies to be matched to specific

learning outcomes. More broadly, institutions of any kind with a strong culture of formative evaluation of learning/training or customer service would not find it difficult to reconcile the cyclical nature of the model with their current practice.

The common factor in all these cases is the alignment of this model within a pre-existing larger framework for quality assurance. The ultimate token of successful application might be that it loses its separate identity as it is absorbed into and contributes to the larger system.

A final area of possible development is the application of the model to a test that is entirely delivered at a distance. The question posed in section 10.2 above was whether genuinely communicative language tests be delivered at a distance over the internet. Visual contact and paralinguistic communication could be provided by video-conferencing supported by multi-media delivery of tasks, similar to the Five Star test event but remotely rather than locally. Such interaction would need to be video-recorded and analysed to compare the type of interaction generated with conventional face-to-face tests. The implications of violation of the local context of communicative testing would need to be examined; is there a sense in which the context could be interpreted as virtual rather than physical? External and consequential validity systems would have to be adapted to elicit electronic feedback. Looking at the validity evidence overall, how communicative could such tests really be said to be? The expert panel, as discussed in earlier chapters, is best operated anyway on the basis on anonymous contributions without physical encounters, and this can equally be done over any distance. In principle, there seems to be an opportunity for developing the model to validate a new generation of internet-based tests of language proficiency.

Bibliography

Adams, R.J. and Siek-Toon Khoo, (1996), *Quest: the interactive test analysis system*, Victoria: The Australian Council for Educational Research

Alderson, J. C. (1990) 'Learner-centred testing through computers: institutional issues in individual assessment', in de Jong and Stevenson (1990) *op. cit.*

Alderson, J. C. and A. Beretta (1992), *Evaluating second language education*, Cambridge University Press

Alderson, J.C. and B. North (eds) (1995), *Language Testing in the 1990s*, Phoenix ELT/ Prentice Hall International

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

Al-Ghamdi, G. A. (1994) 'A model for scoring speaking tests: the King Faisal Air Academy's oral proficiency test', *IATEFL Testing Newsletter*, July 1994

ALTE website <http://www.alte.org>

Anastasi, A. (1982) *Psychological testing*, Macmillan

Andrews, J and R. Fay (1999) 'Interculturalising communicative language testing' in D. Killick and M. Parry (eds) (1999) *Languages for cross-cultural capability: promoting the discipline, making boundaries and crossing borders*, Conference proceedings: Leeds Metropolitan University

Angoff, W.H. (1988) 'Validity: an Evolving Concept' in Wainer and Braun (1988) *op. cit.*

APA (American Psychological Association) (1954) 'Technical Recommendations for Psychological Tests and Diagnostic Techniques', *Psychological Bulletin*, 51 (2) (Supplement)

APA (American Psychological Association) (1966) *Standards for Educational Psychological Tests and Manuals*, Washington DC: American Psychological Association

APA (American Psychological Association) (1974) *Standards for Educational Psychological Tests*, Washington DC: American Psychological Association

APA (American Psychological Association) (1985) *Standards for Educational Psychological Testing*, Washington DC: American Psychological Association

APA (American Psychological Association) (1999) *Standards for Educational and Psychological Testing*, Washington DC: American Psychological Association

- Austin, J.L. (1962), *How to do things with words*, Oxford University Press
- Bachman, L.F. (1990) *Fundamental considerations in language testing*, Oxford University Press
- Bachman, L.F. (2000) 'Modern language testing at the turn of the century: assuring that what we count counts', *Language Testing*, 17:1.
- Bachman, L.F. and A.S. Palmer (1980) A multitrait-multimethod investigation into the construct validity of six tests of speaking and reading', in Palmer, Groot and Tropper (1981) *op. cit.*
- Bachman, L.F. and A.S. Palmer (1996) *Language testing in practice*, Oxford University Press
- Baker, R. (1997) *Classical test theory and item response theory in test analysis* Special Report No 2: Language Testing Update, Centre for Research in Language Education, Lancaster University
- Banerjee, J. (2000), 'Content discussion: the English Language Skills Assessment Testing Programme', paper given at BALEAP Professional Issues Meeting, Nottingham, UK 13 May 2000
- Billows, L. (1961) *The techniques of language teaching*, Longman
- Bloomfield, L (1935) *Language*, Allen & Unwin
- British Council (current, no date) *International English Language Testing Service*
- Brown, H. (1994) *Principles of language learning and teaching*. Prentice Hall Regents, Englewood Cliffs
- Brown, J.D. (1996) *Testing in language programs*, Prentice Hall Regents
- Brumfit C.J. and K. Johnson (eds) (1979) *The communicative approach to language teaching*, Oxford University Press
- Bruton, A. (1999) *Task-based instruction and its closest relations*, paper delivered at IATEFL International Conference, Edinburgh, April 1999
- Busch, M. (1997) 'Messick's Validity', contribution to LTEST-L discussion list 3 November 1997: LTEST-L@LISTS.PSU.EDU
- Campbell, D.T. and D.W. Fiske (1959) 'Convergent and Discriminant Validation by the Multitrait-multimethod matrix', *Psychological Bulletin*, 56, 2
- Canale, M. and M. Swain (1980). 'Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing', *Applied Linguistics*, 1

- Canale, M. and M. Swain (1981). 'A theoretical framework for Communicative Competence', in Palmer, Groot and Trosper (1981) *op. cit.*
- Carroll, B.J. (1980), *Testing communicative performance*, London, Pergamon Press
- Carroll, B.J. and R. West (1989) *ESU Framework: Performance scales for English language examinations*, Longman
- Carroll, J.B. (1973), 'Foreign language testing: will the persistent problems persist?', in Maureen Concannon O'Brien (ed), *ATESOL testing in second language teaching: new dimensions*, Dublin University Press
- CATS (Computer Assisted Training System) website <http://www.cats-training.com>
- Chalhoub-Deville, M (1997) 'Theoretical models, assessment frameworks and test construction', *Language Testing* 14,1
- Chomsky, N. (1959) A review of B.F. Skinner's 'Verbal Behaviour', *Language* 35,1
- Chomsky, N. (1965) *Aspects of the Theory of Syntax*, Boston: MIT Press
- Clark, J.L.D. (1975) 'Theoretical and technical considerations in oral proficiency testing' in Jones R.L. and B. Spolsky (eds) (1975) *Testing language proficiency*, Arlington, Virginia: Center for Applied Linguistics
- Clark, J.L.D. (ed) (1978), *Direct testing of speaking proficiency: theory and application*, Educational Testing Service, Princeton NJ
- Clyne, M. (1996) *Intercultural communication at work: Cultural values in Discourse*, Cambridge University Press
- Cohen, L. and L. Manion (1994) *Research Methods in Education* (4th edition), Routledge
- Commonwealth Office of Education (1965) *Situational English*, Longman
- Coombe C.A. (1998) (ed) *Current Trends in English Language Testing : Conference Proceedings for CTELT 1997 and 1998*. TESOL Arabia, Al-Ain, United Arab Emirates.
- Corcoran, S. and N. Jones (1999), 'Computer adaptive language testing', presentation at IATEFL International Conference, Edinburgh, March 1999
- Crocker, L. and J. Algina (1986) *Introduction to Classical and Modern Test Theory*, Harcourt Brace Jovanovich
- Cronbach, L.J. (1971) 'Validity' in Thorndike, R.L. (1971) *Educational measurement*, Washington DC: American Council on Education

- Cronbach, L.J. (1988) Construct validation after thirty years' in R.L. Linn (ed) *Intelligence: measurement, theory and public policy*, University of Illinois Press
- Cronbach, L.J. and P.E. Meehl, (1972) Construct Validity in Psychological Tests, in V.H. Noll et al, (eds) *Introductory Readings in Educational Measurement*, Boston: Houghton Mifflin
- Cumming, A. and R. Berwick, (eds) (1996). *Validation in language testing*. Multilingual Matters, Clevedon, Avon
- Curran, C. (1976) *Counseling-learning in second languages*, Apple River, Illinois: Apple River Press
- Cziko, G.A. (1983) 'Psychometric and edumetric approaches to language testing' in Oller (1983) *op. cit.*
- Davies, A. (ed) (1968) *Language testing symposium*, Oxford University Press
- Davies, A. (1978) 'Language testing', *Language Teaching and Linguistics Abstracts*, 11, 3/4
- Davies, A. (1990) 'Operationalising uncertainty in language testing', in de Jong and Stevenson (1990) *op. cit.*
- Davies, P., J. Roberts and R. Rossner (1975) *Situational Lesson Plans*, Macmillan
- de Jong, J.H.A.L. and D.K. Stevenson (1990). *Individualising the assessment of language abilities*. Multilingual Matters, Clevedon, Avon
- Delbecq, A.L., A.H. Van de Ven and D.H. Gustafson (1975) *Group techniques for programme planning: a guide to Nominal and Delphi processes*, Scott, Foresman
- Deville, C. (2000) 'Re: CATS', message posted on LTST-L list, 21 February 2000 (LTEST-L@LISTS.PSU.EDU)
- Drew, P. and J. Heritage (eds) (1992), *Talk at work*, Cambridge University Press
- ELI (English Language Institute) (1999), *Examiners Manual for 1999-2000 Examination for the Certificate of Proficiency in English*, Testing and Certification Division, University of Michigan
- ETS (Educational Testing Service) (1993) *TOEIC research summaries number 1: relating TOIEC scores to oral proficiency interview ratings*, Educational Testing Service, Princeton NJ
- ETS (Educational Testing Service) (1998a), *Products and Services Catalog*, Educational Testing Service: Princeton New Jersey
- ETS (Educational Testing Service) (1998b), *Computer-Based TOEFL Score User Guide*, Educational Testing Service: Princeton New Jersey

- ETS (Educational Testing Service) (1999), *Preparing Students for the Computer-Based TOEFL*, Educational Testing Service: Princeton New Jersey
- ETS (Educational Testing Service) (2000), *The Computer-Based TOEFL: enhancing test quality*, Educational Testing Service: Princeton New Jersey
- Faerch, C. and G. Kasper (1983) (eds) *Strategies in interlanguage communication*, Longman
- Ffrench, A. (2000) 'A study of qualitative differences between CPE individual and paired speaking tests' paper given at APPI Conference, Porto, Portugal April 2000
- Firth, J.R. (1957) *Papers in linguistics*, Oxford University Press
- Foot, M.C. (1999a) 'Relaxing in pairs', *ELT Journal* 53,1:36-41
- Foot, M.C. (1999b) 'Reply to Saville and Hargreaves', *ELT Journal* 53,1:52-53
- Frederiksen, J.R. and A. Collins (1989) 'A systems approach to educational testing', *Educational Researcher* 18,9:27-32
- Fulcher, G. (1987) 'Tests of oral performance: the need for data-based criteria', *ELT Journal* 41: 287-291
- Fulcher, G. (1994) 'Some priority areas for research in oral language testing', *Language Testing Update*, 15, Spring 1994 39-47
- Fulcher, G. (1996) 'Testing tasks: issues in task design and the group oral' in *Language Testing* 13,1
- Fries, C.C. (1945) *Teaching and learning English as a foreign language*, Ann Arbor: University of Michigan Press
- Gardner, R.C. and W.E. Lambert (1972) *Attitudes and Motivation in Second Language Learning*, Newbury House
- Gattegno, C. (1976) *The common sense of foreign language teaching* NewYork: Educational Solutions Inc
- Gibson, E. J., P. W. Brewer, A. Dholakia, M. A. Vouk and D. L. Bitzer (1996) 'A Comparative Analysis of Web-Based Testing and Evaluation Systems', on North Carolina State University website at <http://renoir.csc.ncsu.edu/MRA/Reports/WebBasedTesting.html>
- Graham, T. (1997) 'Content review against Trinity College London Spoken English Grade Exams' in Underhill (1997) *op. cit.*
- Grice, H.P. (1975) 'Logic and conversation' in P. Cole and J.L. Morgan (eds) *Speech acts*, New York: Academic Press

- Grice, H.P. (1989), *Studies in the way of words*, Harvard University Press
- Gulinello F. and L. Durso (1998) 'Computer-Assisted EFL Placement for Speaking and Listening Courses', *Language Testing Update* 23, Spring 1998
- Gumperz, J.J. (1982) *Discourse strategies*, Cambridge University Press
- Gumperz, J.J. and D. Hymes (eds) (1970) *Directions in sociolinguistics*, Holt Rinehart & Winston
- Gunnarsson, B. (1978) 'A look at the content similarities between intelligence, achievement, personality, and language tests' in Oller and Perkins (eds) (1978) *op. cit.*
- Halliday, M.A.K, A. McIntosh, and P. Strevens (1964) *The linguistic sciences and language teaching*, Longman
- Harrell, A.T. (1978) *New Methods in Social Science Research*, New York
- He, A.W. (1998) 'Answering Questions in LPIs: A Case Study', in Young and He (1998) *op. cit.*
- He, A.W. and R. Young (1998) 'Language Proficiency Interviews: A Discourse Approach', in Young and He (1998) *op. cit.*
- Heaton, J.B. (1975, new edition 1988) *Writing English language tests*, Longman
- Heaton, J.B. (ed) (1982) *Language testing*, Oxford: Modern English Publications
- Helm, J. (ed) (1968) *Proceedings of the 1967 Spring Meeting of the American Ethnological Society*, University of Washington Press
- Henning, G. (1987) *A Guide to Language Testing*, Newbury House
- Henning, G. (1990) 'National Issues in Individualized Assessment: The Consideration of Specialization Bias in University Language Screening Tests', in de Jong and Stevenson, 1990 *op. cit.*
- Henning, G., T. Hudson and J. Turner (1985) 'Item response theory and the assumption of unidimensionality for language tests', *Language Testing* 2,2:141-154
- Hill, R. A. (1990) *ToPE - The Test of Proficiency in English Manual*, University of Edinburgh
- Hill, R.A. (1995) 'ToPE: Test of Proficiency in English: The Development of an Adaptive Test' in Alderson and North (eds) (1995) *op. cit.*
- Hornby, A.S. (1950) 'The situational approach in language teaching', *English Language Teaching* IV,4

- Howatt, A.P.R. (1984) *A history of English language teaching*, Oxford University Press
- Hudson L, (1967) *Contrary imaginations*, Penguin
- Hughes, A. (1989) *Testing for language teachers*, Cambridge University Press
- Hughes, A. (1990) 'Response to Spolsky' in de Jong and Stevenson (1990) *op. cit.*
- Huhta, A. and E. Randell (1996) 'Multiple-choice summary: a measure of text comprehension', in Cumming and Berwick (1996) *op. cit.*
- Hutchinson, T. (1991) *Introduction to project work*, Oxford University Press
- Hutchinson, T. and A. Waters (1987) *English for specific purposes: a learner centred approach*, Cambridge University Press
- Hymes, D. (1968) 'Linguistic problems in defining the concept of the tribe', in Helm (1968) *op. cit.*
- Hymes, D. (1970) 'On communicative competence', in Gumperz and Hymes (1970) *op. cit.*
- Jenkins, J. (1997) 'Testing Pronunciation in Communicative Exams', *Speak Out!* IATEFL Pronunciation SIG Newsletter 20
- Johnson, M. and A. Tyler (1998) 'Re-analyzing the OPI: How Much Does It Look like Natural Conversation?', in Young and He (1998) *op. cit.*
- Jones, N. (1991) 'Test Item Banker: An item Bank for a Very Small Micro', in Alderson and North (1995) *op. cit.*
- Joos, M. (ed) (1957) *Readings in linguistics I*, University of Chicago Press
- Katona, L. (1998) 'Meaning Negotiation in the Hungarian Oral Proficiency Examination of English', in Young and He (1998) *op. cit.*
- Kendall, M.G. and W.R. Buckland (1982) *A dictionary of statistical terms*, Longman
- Kline, P. (1993) *The Handbook of Psychological Testing*, Routledge
- Koike, D.A. (1998) 'What Happens When there's No One to Talk to? Spanish Foreign Language Discourse in Simulated Oral Proficiency Interviews', in Young and He (1998) *op. cit.*
- Kontoulis, E. (1997) 'Content review against IELTS test' in Underhill (1997) *op. cit.*
- Krashen, S. (1981) *Second language acquisition and second language learning*, Oxford, Pergamon Press

- Krashen, S. (1982) *Principles and practices in second language acquisition*, Oxford, Pergamon Press
- Krashen, S. and T.D. Terrell (1983) *The natural approach: language acquisition in the classroom*, Oxford, Pergamon Press
- Labov, W. (1966) *The social stratification of English in New York City*, Center for Applied Linguistics
- Lado, R. (1957) *Linguistics across cultures: applied linguistics for language teachers* Ann Arbor: University of Michigan Press
- Lado, R. (1961) *Language testing*, Longman
- Lado, R. (1964) *Language teaching, a scientific approach*, McGraw Hill
- Lantolf, J.P. and W. Frawley (1985) 'Oral proficiency testing: a critical analysis', *Modern Language Journal* 69: 337-345
- Laurier, M. (1996) 'Using the information curve to assess CAT efficiency', in Cumming and Berwick (1996) *op. cit.*
- Lazaraton, A. (1992) 'The structural organization of a language interview: a conversation analytic perspective' *System* 20,3
- Lazaraton, A. (1996) 'Interlocutor support in oral proficiency interviews: the case of CASE', *Language Testing* 13,2
- LCCI (London Chamber of Commerce and Industry) (current), *English Language Skills Assessment*, LCCI Examinations Board: London
- Lee, Y.P., A.C.Y. Fok, R. Lord, and G. Low (eds) (1985), *New directions in language testing*, Pergamon
- Lewkowicz, J.A. (2000) 'Authenticity in language testing: some outstanding questions', *Language Testing* 17,1
- Linn, R.L., E.L. Baker and S.B. Dunbar (1991) 'Complex, performance-based assessment: expectations and validation criteria', *Educational Researcher* 20,8:15-21
- Linstone, H.A. and M. Turoff (eds) (1975) *The Delphi method: techniques and applications*, Addison-Wesley
- Lozanov, G. (1978) *Suggestology and the outlines of suggestopedy*, New York: Gordon and Breach
- Lumley, T. (1993). "The notion of subskills in reading comprehension tests: an EAP example" *Language Testing*, volume 10 number 2, 1993
- Lynch, B. (1996), *Language program evaluation*, Cambridge University Press

- Mackay, R. (1993), "Programme evaluation as -a management tool for both accountability and improvement" *IATEFL ELT Management Newsletter* 12, June 1993
- Madsen, H.S. and R.L. Jones (1981) 'Classification of Oral Proficiency Tests', in Palmer, Groot and Trosper (1981) *op. cit.*
- Malabonga, V. (1998) 'The Computerised Oral Proficiency Interview', *Language Testing Update* 24 Autumn 1998 Centre for Research in Language Education: Lancaster
- Malinowski, B. (1923) 'The problem of meaning in primitive languages' in C.K. Ogden and I.A. Richards (1923) *The meaning of meaning*, Routledge & Kegan Paul
- Masters, G.M. (1982) 'A Rasch model for partial credit scoring', *Psychometrika*, 47,2
- McCarthy, A. (1997) 'Content review against UCLES EFL exams' in Underhill (1997) *op. cit.*
- McCarthy, M. (1991) *Discourse analysis for language teachers*, Cambridge University Press
- McNamara, T.F. (1996), *Measuring second language performance*, Longman
- Messick, S. (1981) 'Constructs and their vicissitudes in educational and psychological measurement', *Psychological Bulletin* 89
- Messick, S. (1988) 'The once and future issues of validity: assessing the meaning and consequences of measurement' in Wainer and Braun (1988) *op. cit.*
- Messick, S. (1989) 'Validity', in R.L. Linn (ed) (1989) *Educational measurement*, Macmillan/American Council on Education
- Messick, S. (1994), 'The interplay of evidence and consequences in the validation of performance assessments', *Educational Researcher* 23,2
- Milanovic, M., N. Saville, A. Pollitt and A. Cook (1996). 'Developing rating scales for CASE: theoretical concerns and analyses' in Cumming and Berwick (1996) *op. cit.*
- Moder, C.L. and G.B. Halleck (1998) 'Framing the Language Proficiency Interview as a Speech Event: Native and Non-Native Speakers' Questions', in Young and He (1998) *op. cit.*
- Morrow, K. (1977) *Techniques of evaluation of a notional syllabus*, Royal Society of Arts, mimeo
- Morrow, K. (1979) 'Communicative language testing: revolution or evolution?' in Brumfit and Johnson (1979) *op. cit.*
- Moskowitz, G (1978) *Caring and sharing in the foreign language class*, Rowley, Mass: Newbury

Moss, P. (1992) 'Shifting Conceptions of Validity in Educational Measurement: Implications for Performance Assessment', *Review of Educational Research* 62, 229-258

Munby, J. (1978) *Communicative syllabus design*, Cambridge University Press

Netten, G. (2000a) 'TOEFL's hopes for CBT in the next millennium', *English Language Gazette*, January 2000

Netten, G. (2000b) 'TOEFL and computer-based TOEFL', paper given at BALEAP Professional Issues Meeting, Nottingham, UK 13 May 2000

North, B. and G. Schneider (1998), 'Scaling descriptors for language proficiency scales', *Language testing*, 15:2

Nunan, D. (1992) *Research methods in language learning*, Cambridge University Press

Ockenden, M. (1972) *Situational Dialogues*, Longman

Oller, J.W. (1979) *Language tests at school*, Longman

Oller, J.W. (ed) (1983) *Issues in language testing research*, Newbury House

Oller, J.W. and K. Perkins (eds) (1978) *Language in education: testing the tests*, Newbury House

Ordinate (1998), *The PhonePass Test*, published by Ordinate: Menlo Park, California

Ordinate (1999), *Validation Summary for PhonePass*, published by Ordinate: Menlo Park, California

Ordinate website <http://www.ordinate.com>

Palmer, A.S., P.J.M. Groot and G.A. Troster (eds) (1981) *The Construct Validation of Tests of Communicative Competence*, Washington D.C. : TESOL

Palmer, A.S. and P.J.M. Groot (1981) 'Introduction' to Palmer, Groot and Troster (1981) *op. cit.*

Palmer, H.E. (1922) *The principles of language study*, Harrap

Palmer, H.E. (1932) *This language learning business*, Harrap

Parker, R.J. (1997) 'Appendix VIII: a framework for training and licensing Five Star assessors' in Underhill (1997) *op. cit.*

Pattison, B. (1964) 'Modern methods of language teaching', *English Language Teaching* 19,1

- Pollard, J.D.E. (1994) 'Proficiency testing', *IATEFL Testing Newsletter*, July 1994
- Pollard J.D.E. (1997) 'Saudi Development and Training's Five Star Proficiency Test Project'. In Coombe C.A. (1998) *op. cit.*
- Pollard J.D.E. (1998a) 'The Influence of Assessor Training on Rater-as-Interlocutor Behaviour During a Computer-Resourced Oral Proficiency Interview-cum-Discussion (OPI/D) known as the Five Star Test'. In Coombe C.A. (1998) *op. cit.*
- Pollard J.D.E. (1998b) 'Research and development: a complex relationship – Part I' *Language Testing Update*, issue 24, autumn 1998, Centre for Research in Language Education, Lancaster University
- Pollard J.D.E. (1999) 'Research and development: a complex relationship – Part II' *Language Testing Update*, issue 26, autumn 1999, Centre for Research in Language Education, Lancaster University
- Pollard, J.D.E. (work in progress) unpublished PhD dissertation, University of Surrey
- Pollard, J.D.E and N. Underhill (1996) 'Developing and researching validity for a computer-resourced proficiency interview test', *Language Testing Update*, issue 20, autumn 1996, Centre for Research in Language Education, Lancaster University
- Quade, E.S. (1977), *Analysis for public decisions*, Elsevier: New York
- Rasch, G. (1960) *Probabilistic models for some intelligence and attainment tests*, Copenhagen: Danish Pedagogical Institute
- Rea-Dickins, P. and K. Germaine (1992), *Evaluation*, Oxford University Press
- Resources in Language Testing website <http://www.surrey.ac.uk/ELI/ltr.html>
- Richards, J.C. and T.S. Rodgers (1986) *Approaches and methods in second language teaching*, CUP
- Riggenbach, H. (1998) 'Evaluating Learner Interactional Skills: Conversation at the Micro Level', in Young and He (1998) *op. cit.*
- Robinson, R. (1996) *Understanding the Saudi labour market*, Saudi Development and Training, Dammam, Saudi Arabia
- Ross, S. (1998) 'Divergent Frame Interpretations in Oral Proficiency Interview Interaction', in Young and He (1998) *op. cit.*
- Ross, S. and R. Berwick (1992) 'Accommodative questions in oral proficiency interviews', *Language Testing*, 9
- Sacks, H., E.A.Schegloff and G. Jefferson, (1974), 'A simplest systematics for the organization of turn-taking in conversation', *Language*, 50

- Saville, N. and P. Hargreaves (1999) 'Assessing Speaking in the Revised FCE', *ELT Journal* 53,1:42-51
- Scovel, T. (1979) Review of 'Suggestology and the outlines of suggestopedy', *TESOL Quarterly*, 13
- SDT (no date) *The Five Star English Test Assessor's Manual*, Saudi Development and Training
- Searle, J.R. (1969) *Speech acts: an essay in the philosophy of language*, Cambridge University Press
- Searle, J.R. (1979) *Expression and meaning: studies in the theory of speech acts*, Cambridge University Press
- Shohamy, E. (1994) 'The validity of direct versus semi-direct oral tests', *Language Testing* 11,2
- Skehan, P. (1984) 'Issues in the testing of English for Specific Purposes', *Language Testing* 1: 202-220
- Skehan, P. (1989) *Individual differences in second-language learning*, Edward Arnold
- Spolsky, B. (1975) 'Language testing: art or science?' Paper delivered 27.8.1975 at 4th AILA World Congress in Stuttgart, Germany
- Spolsky, B. (1985) 'The limits of authenticity in language testing', *Language Testing* 2,1: 31-40
- Spolsky, B (1990) 'Social aspects of individual assessment' in de Jong and Stevenson (1990) *op. cit.*
- Stansfield, C.W. and D.M. Kenyon, (1996) 'Comparing the scaling of speaking tasks by language teachers and the ACTFL guidelines', in Cumming and Berwick (1996) *op. cit.*
- Stevenson, D.K. (1981) Beyond Faith and Face Validity: The Multitrait-Multimethod Matrix and the Convergent and Discriminant validity of Oral Proficiency Tests' in Palmer, Groot and Trosper (1981) *op. cit.*
- Stevenson, D.K. (1985a) 'Authenticity, validity and a tea party', *Language Testing*, 2,1: 41-47
- Stevenson, D.K. (1985b) 'Pop validity and performance testing', in Lee et al (1985) *op. cit.*
- Stevick, E.W. (1976) *Memory, meaning and method*, Rowley, Mass: Newbury House
- Stevick, E.W. (1980) *Teaching languages: a way and ways*, Rowley, Mass: Newbury House

Stevick, E.W. (1982) *Teaching and learning languages*, Cambridge University Press

Streiff, V. (1978) 'Relationships among oral and written cloze scores and achievement test scores in a bilingual setting' in Oller and Perkins (eds) (1978) *op. cit.*

Tarone, E. (1980) 'Communication strategies, foreigner talk, and repair in interlanguage', *Language Learning*, 30

TOEFL website <http://www.toefl.org>

Trinity College Grade Examinations in Spoken English (current), Trinity College London

UCLES (University of Cambridge Local Examinations Syndicate) (1995) *Certificates in Communicative Skills In English Handbook*, University of Cambridge

UCLES (current) Handbooks for 'Main suite' and additional English language proficiency examinations: Key English Test, Preliminary English Test, First Certificate in English, Certificate in Advanced English, Certificate of Proficiency in English, CommuniCAT, Cambridge (see also under *UCLES* in Glossary)

UCLES EFL website <http://www.cambridge-eefl.org.uk>

Underhill, N. (ed) (1997) *Final report on the critical review of the Five Star test*, unpublished consultancy report for Saudi Development and Training, Sheffield Hallam University

Underhill, N. (1987) *Testing spoken language*, Cambridge University Press

Underhill, N. (1982) 'The great reliability/validity trade-off' in Heaton (1982) *op. cit.*

Upshur, J.A. and C. E. Turner (1999), 'Systematic effects in the rating of second-language speaking ability: test method and learner discourse', *Language Testing* 16,1

Valette, R.M. (1967, second edition 1977), *Modern Language Testing* Harcourt Brace Jovanovich

Van Lier, L. (1989) 'Reeling, writhing, drawling, stretching, and fainting in coils: oral proficiency interviews as conversation', *TESOL Quarterly*, 23,3 September 1989

Van Lier, L. (1996) *Interaction in the Language Curriculum*, Longman

Wainer, H. and H.I. Braun (eds), (1988) *Test validity*, Hillsdale NJ, Lawrence Erlbaum Associates

Wainer H., N.J. Dorans, R. Flaugher, B.F. Green, R.J. Mislevy, L. Steinberg and D. Thissen (1990) *Computerized adaptive testing: A primer* Hillsdale, NJ: Erlbaum

Weir, C.J. (1990), *Communicative Language Testing*, Prentice Hall

- Weir C.J. and Roberts J. (1994) *Evaluation in ELT*, Blackwell
- Weiss, D.J. (1982) Improving Measurement Quality and Efficiency with Adaptive Tests, *Applied Psychological Measurement*, 6,4:473-492
- Wellesley, S. (1993), "Project based review in practice" in *IATEFL ELT Management Newsletter* 13, October 1993
- Widdowson, H. (1979) *Explorations in applied linguistics*, Oxford University Press
- Wierzbicka, A. (1985) 'Different Cultures, Different Languages, Different Speech Acts', *Journal of Pragmatics* 9, 145-178
- Wijgh, I.F. (1993) 'The Art of Getting Consensus/manufacturing Consensus validation of a Theoretical Framework for Oral Interaction by Expert Judgment', unpublished article based on PhD thesis, CITO, Netherlands
- Wilkins, D.A. (1972) *An investigation into the linguistic and situational common core in a Unit/Credit system*, Council of Europe
- Wilkins, D.A. (1976), *Notional Syllabuses*, Oxford University Press
- Wilkins, D.A. (1979) 'Grammatical, situational and notional syllabuses' in Brumfit and Johnson (1979) *op. cit.*
- Williams, S. (2000) 'Multi-lingual computer-based language assessment', presentation at IATEFL International Conference, Dublin, March 2000
- Willis, J. (1996) *A framework for task-based learning*, Longman
- Wilson, K. (1993) *Relating TOEIC scores to oral proficiency interview ratings*, TOEIC Research Summaries no 1, Educational Testing Service: Princeton, New Jersey
- Wright, B.D. and G.N. Masters, (1982), *Rating scale analysis*, Chicago: MESA Press
- Wright, B.D. and M.H. Stone (1979) *Best test design*, Chicago: MESA Press
- Yoshida-Morise, Y. (1998) 'The Use of Communication Strategies in Language Proficiency Interviews', in Young and He (1998) *op. cit.*
- Young, R. and A. W. He (eds) (1998) *Talking and testing: discourse approaches to the assessment of oral proficiency* (Studies in Bilingualism Volume 14) John Benjamins Publishing Company: Amsterdam and Philadelphia
- Zumbo, B.D. (1999) *A handbook on the theory and methods of differential item functioning*, Ottawa (ON): Directorate of Human Resources Research and Evaluation, Department of National Defense

Appendices

Appendix	Page
Appendix I Glossary and abbreviations.....	434
Appendix II Sample task cards from Five Star test.....	436
Appendix III Sample page from Five Star test algorithm.....	440
Appendix IV Criterion validation against external tests	441
Appendix V A framework for training & licensing Five Star assessors	448
Appendix VI Sample of Five Star test outputs: 30 candidates on 28 tasks	450

Adaptive An adaptive test is one in which responses to earlier items influence the selection of later items in each test administration, so that two consecutive subjects taking the same test actually face few or none of the same tasks or items.

APA, the American Psychological Association, [publishers of a set of Standards for Educational and Psychological Testing, which have been revised and republished approximately every ten years from 1954 to 1999]

“**Assessment** is a process of gathering information to meet a broad range of evaluation needs, and differs from *testing* in that it uses multiple indicators and sources of evidence. By definition, therefore, an assessment program should employ some variety of strategies and procedures for observing, collecting, and evaluating student work and student learning.

The term **alternative assessment** is used to distinguish ... new kinds of assessments from conventional, primarily multiple choice, paper-and-pencil tests. One well-known example of alternative assessment is the *portfolio*, a purposeful collection of student work overtime in a particular subject area. ...

Naturalistic assessment refers to evaluation that is rooted in the natural setting of the classroom and involves observation of student performance and behavior in an informal context. Documentation, a naturalistic method, is a process of classroom observation and record keeping over time, across learning modalities, and in coordination with colleagues” *Focus: capturing the power of classroom assessment*, ETS, 1995

Assessor, see under **interviewer**

Backwash, the impact of test design on teaching methodology; in contrast to ‘bow-wave’, the reflection in testing methods on changes in teaching methodology

C-test, a gapfill test where the first letter or letters of the missing word are provided as clues

CAE and CPE, see **UCLES**

CAT, computer-adaptive test

Candidate, the person taking the test

Computer-based test, a form of assessment where the task is delivered by computer. It may or may not be mediated by a live interlocutor, and the candidate’s response may be assessed directly by the computer, or by the interlocutor, or a combination of both

CRT, criterion-referenced tests, see section 3.3.4

Direct, semi-direct, indirect A **direct** test (also known as OPI, oral proficiency interview, in USA). is one that involves face-to-face oral interaction between learner and interviewer or interlocutor. A **semi-direct** test elicits spoken language from the learner, but it is recorded on audio or video for subsequent rating (eg ARELS, TSE) and there is no interaction, in the everyday sense (also known as SOPI, semi-direct (or ‘simulated’) oral proficiency interview, in USA). An **indirect** test does not elicit spoken language at all, but claims validity on the basis of statistical correlation with direct tests. (this terminology originated in Clark 1975)

EFL, TEFL, (teaching) English as a Foreign Language

ETS The Educational Testing Service, Princeton, New Jersey, USA, who publish the TOEFL, TOEIC Test of Spoken English (TSE) and Test of Written English (TWE). TOEFL and TOEIC contain no direct speaking component, but may be supplemented by the TSE, which is a semi-direct oral test where candidate answers are recorded on tape - see **direct**

IELTS International English Language Testing System, published by UCLES, jointly managed by UCLES, the British Council, and IDP Education Australia (see content comparison in appendix IV).

Indirect test, see **direct**

Interviewer, interlocutor “An **interviewer** is a person who talks to a learner in an oral test and ... also takes the role of assessor

An **interlocutor** is a person who talks with a learner in an oral test, but who is not required to assess him/her, and whose specific aim is to encourage the learner to display to the assessor his/her oral fluency in the best way possible

An **assessor** is a person who listens to a learner speaking in an oral test and makes an evaluative judgement on what s/he hears.

Marker/rater is someone who is not present at the test itself but later awards marks to the learner on the basis of an audio or video tape recording”. (Underhill 1987)

IRT, item-response theory, a statistical approach to the analysis of test data, see section 3.5

MTMM, the multitrait-multimethod approach to validation, see section 3.2.3

NRT, norm-referenced tests, see section 3.3.4

LPI, Language Proficiency Interview, a widely known and used direct oral interview format, first developed by the Foreign Service Institute of the U.S. Department of State

Naturalistic assessment, see **assessment**

RSA, The Royal Society of Arts Examinations Board first produced the Test of the Communicative Use of English as Foreign Language (CUEFL, later CCSE) in the early 1980s, and was subsequently merged with the University of Cambridge Local Examinations Syndicate (**UCLES**) Cambridge Examinations Board

Semi-direct test, see **direct**

SLA second language acquisition

SOPI see **semi-direct** under **direct test**

TEFL, see **EFL**

TLU, the domain of target language use

TOEFL Test of English as a Foreign Language, see **ETS**

TOEIC Test of English for International Communication, see **ETS**

TSE Test of Spoken English, see **ETS**

TWE Test of Written English, see **ETS**

UCLES The University of Cambridge Local Examination Syndicate is the single biggest examining body in UK for English for speakers of other languages. They produce among other tests a 'main suite' of examinations at five different levels (see content comparison in appendix IV), all known by their initials (Key English Test = KET; Preliminary English Test = PET; First Certificate in English = FCE; Certificate in Advanced English = CAE; Certificate of Proficiency in English = CPE). All the main suite exams contain various subtests, including a direct test of spoken English. UCLES also publishes the Certificates in Communicative Skills in English (CCSE) and CommuniCAT, a computer-adaptive testing service provided under license, and collaborate with the British Council and International Development Program of Education Australia (IDPEA) in the IELTS test. All of these include a direct speaking component.

ame

e the candidate tell his names. Get him explain a little about names if possible. the magnifying s prompts if ssary.

e his names and ID ber into the opriate boxes. Use mouse to set the or in a box. Do not s return se the keyboard w keys.

What's your name?
Is that your first name?



Have you any other names?
Is -- your second name?
Is that your father's name?



What are your other names?
Is that your family name?



What is your last name?



ID Number (max 6 digits)



enter next sequential ID

Comprehension / Pronunciation



Hesitant

Complete

Expansive



4

Appendix II

Sample task cards from Five Star test (continued)

Intermediate Numeracy



e candidate to point h number as he it spoken.

930

72 1/4

6 128

3 913



Ask the candidate to read these calculations.

14 - 3 = 11

5 x 12 = 60

24 / 6 = 4



Ask the candidate to read the numbers listed below in any order.

182

909

9,000

1/2

2,000,100

128

1,515

77

Accuracy

<25%

around 50%

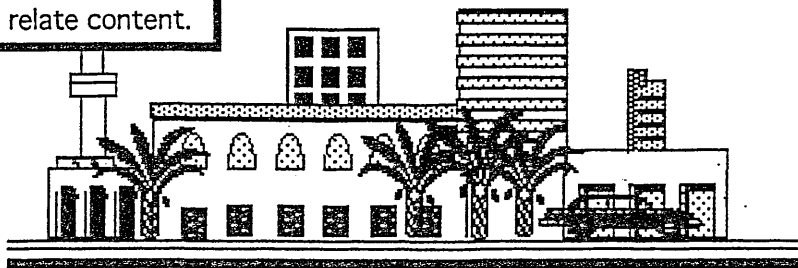
>75%



11

ask candidate to listen carefully. Candidate may listen twice, but no extra assistance should be given.

have candidate relate content.

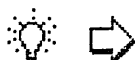


Comprehension

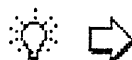
No duality



Duality / no contrast



Contrast



13

Appendix II

Sample task cards from Five Star test (continued)

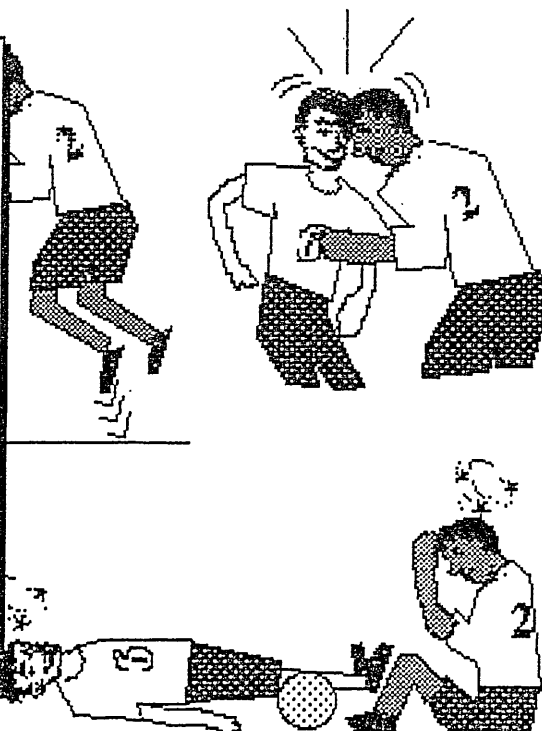
The Footballers

Task 1. The first part of this task tests the ability to explain instructions.

The candidate will hear the following explanation in Arabic: "A series of pictures will appear on the screen one after the other. These pictures portray an incident. Watch the pictures and then describe what happened".

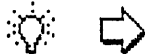
Click the Arabic instruction icon and have the candidate explain the task.

Task 2. Activate the sequence by clicking the "show" button and have the candidate relate the incident. The sequence may be repeated 3 times. On the third showing the interviewer may prompt with questions.

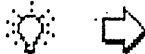


Explanation / Narration

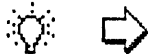
Adequate



High accuracy



24



candidate that one of the texts (1 to 6) is the written message on the road sign in the picture.

He will have 5 seconds to silently skim-read each text and select the best choice. This sequence may be repeated once.

To gain the 'maximum' score, the candidate's correct selection should be made at this point.

The candidate who makes a correct selection only after the repeated sequence and reviewing one or more individual text will gain an 'average' score.

Click the assist button to activate individual reviews.

show 1

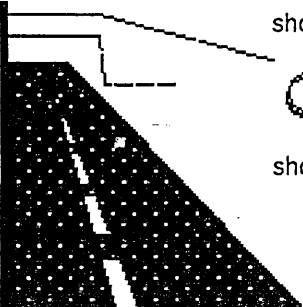
show 2

show 3

show 4

show 5

show 6



1	2	3	4	5	6
Warning contains ous ances. t well from	Warning Heavy goods vehicles are not allowed on this section	Caution The outer case of this appliance should only be opened by an authorized electrician			

Text identification

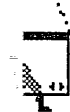


Average

Maximum

assist





appear which make up a paragraph about climactic changes.

The first sentence is given but the remaining 5 are in the wrong order. The candidate's task is to read them and suggest the correct order. Advise the candidate that he will be timed as he does this.

Click 'show' button to reveal the sentences and start timer. As the candidate suggests the correct order write the numbers 2 to 6 in the small boxes to the left by clicking on them with the mouse pointer and typing the number.

When the candidate finishes click the 'timer' button. This reveals the correct order and the time taken.

Accuracy (120 secs)

over 120 secs

4 or 5 correct

100%

rd 89



Climactic Changes



1 There has been a great deal of discussion recently about changes in global climates.

Nor did rainfall vary from its normal pattern over the last 10 years.

However, the local climate here in Saudi Arabia does not seem to have been affected.

A major theme of this discussion suggests that these are caused by pollution from factories, car exhausts and air conditioners.

Last year, for example, coastal temperatures here did not exceed the usual maximum of 42 °C.

Both the amount and monthly distribution follow the rainfall patterns recorded since 1983.

Accuracy (120 secs)

over 120 secs

4 or 5 correct

100%

rd 89

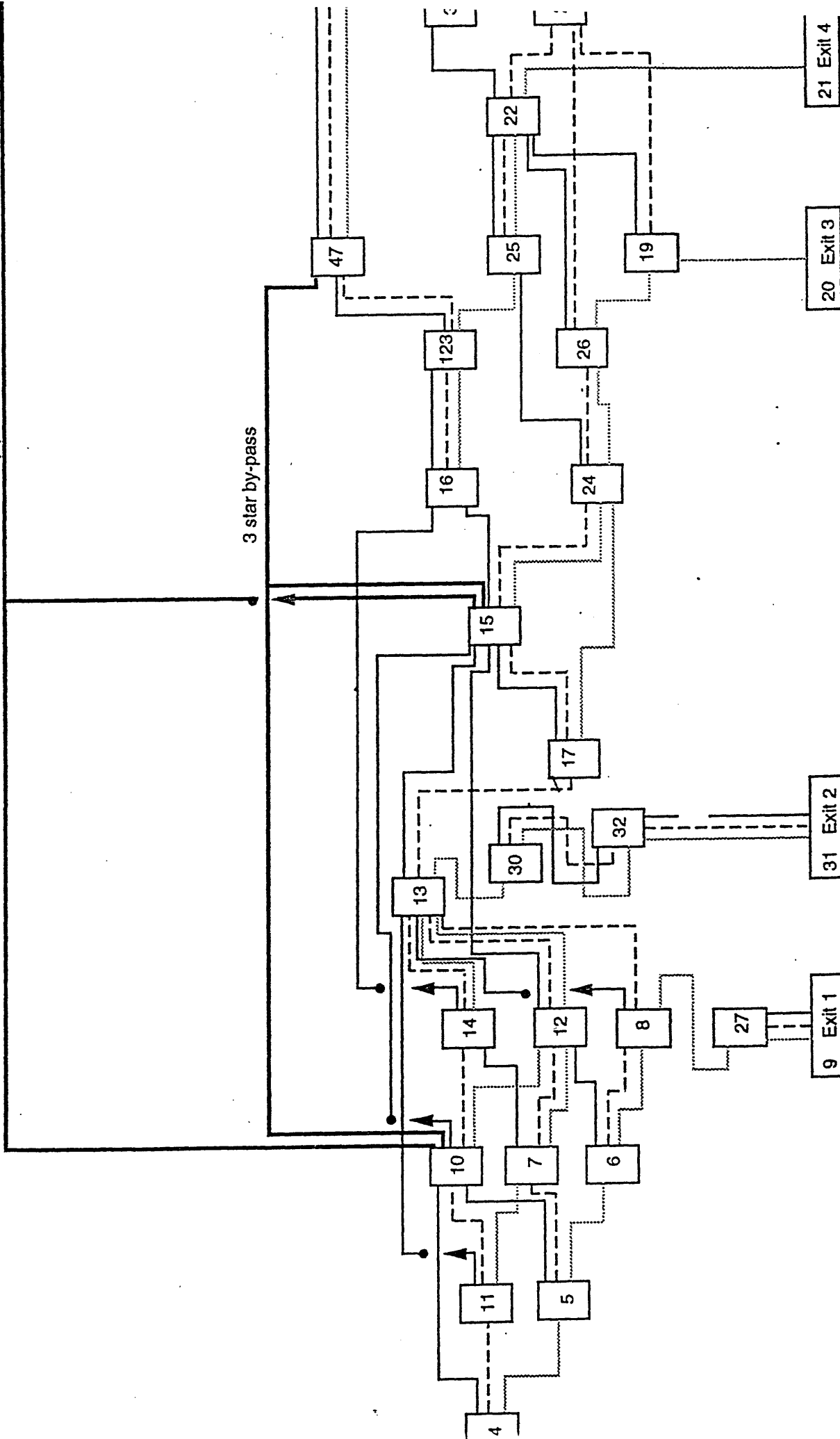


4 star by-pass

3 star by-pass

440

Appendix III - sample page from algorithm



Appendix IV Criterion validation against external tests

This appendix consists of the three content comparisons against other established tests that provided evidence for the criterion and content validation of the Five Star test. These content reviews were written by three different members of the 'expert panel' after they had conducted the other panel activities described in 5.1.1 and 5.1.2 and formed part of the critical review of the Five Star test. They compare the Five Star test with a) the UCLES 'main suite' EFL exams (author Angela McCarthy); b) the Trinity College London Spoken English Grade Exams (author Tim Graham); and c) the IELTS test (author Eileen Kontoulis). The authors are all current or former accredited examiners of the criterion test they are describing.

a) Content review against UCLES 'main suite' EFL exams (Angela McCarthy)

Range of skills

The UCLES exams have separate test components for speaking, listening, writing and reading. The first three are tested very much in isolation, not relying more than absolutely necessary on other productive skills to facilitate the tasks. For example, in listening tests: non-linguistic responses to questions at lower levels, multiple-choice responses, no speaking is involved and written responses are marked for meaning rather than accuracy. In addition, instructions are written as well as spoken.

The Five Star test differs in several ways; the most striking difference throughout is the intensive oral/aural interaction, through which all the skills are tested. This would seem to make whatever task is being explained all the more stressful, and often the candidate has to rely on listening skills to enable him to know what is required. In addition, task performance is judged on the candidate's verbal answer, which makes speaking skills an issue in most of the tasks.

Writing in any communicative or functional sense is currently missing from the Five Star test; it only features at a graphic level. In UCLES exams writing is as important as the other skills.

Levels

UCLES exams range from an elementary exam to an advanced level. The Five Star test is a diagnostic test which makes it different in nature and this accounts for many of the differences in format and testing techniques. The levels which it covers are similar.

Testing techniques

In listening and reading tasks, the tasks are often presented after listening/reading, which tests memory rather than comprehension; although candidates are invited to listen again and told it is not a test of memory, they often seem discouraged at this point. Listening and reading texts do not seem particularly authentic, rather like a text being read aloud, or written to demonstrate a grammar point, which does not represent normal language use. UCLES adapts most listening and reading texts from authentic sources. In speaking and writing tasks, candidates are usually given a role, thereby making the task relevant to a certain context.

Speaking is sometimes tested using picture prompts, sometimes with discussion prompts, or tasks drawing on candidate's personal experience, or knowledge or opinions; these techniques are all very similar in the UCLES exams. There is generally more reliance on non-verbal prompts or stimuli in the latter. The Five Star uses the candidate's first language as input for tasks, where UCLES would of necessity rely on target language input in one form or another, or non-verbal stimuli.

The Five Star test also asks candidates to read aloud. One can only presume that this is testing the relationship between spelling and pronunciation, which is not tested in UCLES exams.

Pronunciation is one of the criteria used for grading in speaking, but not in relation to the written form.

Texts with missing words test vocabulary and reading skills, which is similar to some text types in UCLES. Choosing the best text from a selection as suitable for a particular context is also similar.

The testing of study skills does not form part of UCLES exams in the way the Five Star test uses it - using charts, diagrams and tables from which to select, insert or use information. The reformulating and application of information applies in the very task-oriented writing papers, but there is nothing like this in the other components.

At times, tasks which should have generated more extended language in the Five Star did not e.g. comparing two places the candidate knows; this could have generated a lot of language from the candidate, but usually did not do more than elicit grades for the places. The onus could have been left to the candidate here to take the initiative and direct the discussion, as would happen in an UCLES exam.

Topics

Topics in UCLES exams depend to a certain extent on the level; common areas are: transport, travel and tourism, entertainment, education, occupations, family relationships, the environment, services, fashion, crime, the media and the weather. Many of these came up in the Five Star test too.

Elicitation

The onus is on the interviewer to elicit most language in the Five Star; there is scope for candidate initiative, but the format is very much interviewer controlled and at times the candidate was not allowed the opportunity to expand on a topic - the examiner seemed to talk more and use more interaction strategies than the candidate. Many of the questions which were used to elicit responses were closed questions, with the examiner then left to repair communication breakdown and keep the interaction going. More open questions are needed as well as more time for the candidate to think, react and respond.

Language samples

Connected to the above, there was very little extended speech, except with fairly advanced candidates. Sometimes this is as a result of the questions asked; sometimes as a result of the topics which did not always lend themselves to expansion and occasionally because the control in the tasks is never handed over to the candidate; who is seen very much as responsive in the interaction rather than proactive.

Interaction patterns

The Five Star test is very examiner intensive; all tests are performed in oral/aural mode, with the examiner required to explain most tasks, keep the candidate talking and assess performance. By comparison, only the speaking component of UCLES exams is conducted like this; all but one of the suite of exams are now conducted in paired format, two candidates with two examiners. Research has suggested that the paired format produces interaction which would not be forthcoming from an individual to an examiner; however a part of the interview is conducted in interviewer - candidate interaction as this mode too is beneficial for candidates. The advantage of having two examiners is that one examiner acts as interlocutor, managing and directing the interaction, but retreating at key points to allow candidates to manage it on their own; the other assesses. Assessment appears to be more accurate, carried out in this way. The examiner in the Five Star test is under pressure to help the individual candidate at all stages in the test, as well as assess candidate performance.

Examiner roles

UCLES is moving towards an entirely scripted interview, with very tight timing; the examiner input is limited, just enough to set up an activity, and then withdraw. Intervention after that is only if the candidates are obviously in real difficulty. All of this is in the interests of standardising both format and assessment.

Assessment

Assessment in UCLES exams has become highly standardised in recent years, particularly in speaking components in which variation in assessment is most likely to occur. Assessment in the Five Star test is not very transparent; the criteria which I imagine to be operating are as follows (in order of importance): task achievement; accuracy; range of expression; pronunciation; interaction skills, appropriacy, organisation and cohesion.

b) Content review against Trinity College London Spoken English Grade Exams (Tim Graham)

Format

The format of the two tests is broadly similar since both are primarily oral. Both, too, contain listening tasks, though in Trinity's case these are confined to the advanced stage (Grades 10, 11 and 12). The type of texts used for these advanced stage grades are not dissimilar to the listening texts that would be at roughly level 7, 8 and 9 respectively in terms of difficulty, lexical density and T-unit count. Trinity does not discretely measure listening comprehension below grade 10.

The Five Star test measures reading level and study skills. Trinity does not discretely measure study skills. Trinity does, though, measure reading indirectly through discussion of prepared texts (usually short novels or factual longer texts, such as biographies). Candidates do not read anything during the exam itself, however.

Trinity grades 1 to 3 are really only applicable to young learners, whilst the lower levels of the Five Star are designed for adult learners. From Trinity grade 4 upwards there is discussion of a prepared topic. The sophistication of the topic itself, in terms of content, and discussion increases grade by grade. These topics are selected beforehand by the candidate and in this the discussion will be more rehearsed than that which is elicited by the task pages in the Five Star Test.

The timing of the Trinity grades is set, with 5 minute conversations at the Initial Stage (Grades 1 - 3), 10 minute conversations at the Elementary Stage (Grades 4 - 6), 15 minute conversations at the Intermediate Stage (Grades 7 - 9) and a maximum of 25 minute conversations at the Advanced Stage (Grades 10 - 12). The timing of the Five Star Test would appear to be much more flexible, irrespective of the level of the candidate. In some of the video excerpts of the Five Star Test, candidates at what would be lower levels for Trinity are examined for a good deal longer than would be the case in the grade exams. In terms of discrete skills analysis this would appear to provide the examiner with greater insights, but in terms of overall language proficiency and a judgement as to competence and level, the difference would not seem to be justified by the mismatch in timing as it is possible to make an assessment as to general level from the video extracts in roughly the same time as would be taken by the Trinity Grade exams.

Assessment levels

The Five Star test is a general proficiency measuring examination. In this it differs from Trinity Grade Exams in that the latter are banded tests at specific levels. Candidates are entered at the grade thought by teachers to be most appropriate to their existing level of English language. As long as an appropriate judgement is made on the part of the teachers concerned it should be a matter of the candidate performing to their perceived norm of language use and passing the grade accordingly. With the Five Star, the assessment given is dependent upon the candidate's overall performance against a series of criteria for general language competence. In this regard the two tests differ and, though they are measuring the same kind of proficiency, their approach to this is fundamentally distinct.

Elicitation Techniques

Most of the elicitation for the Trinity Grade Exams is done directly by the examiner. There is usually no intervening vehicle as with the Five Star test via the task pages. The interaction that follows on from the initial identifier page and deals with the pages on home/school and interests in the Five Star test is very similar to the interaction in the early phases of the Trinity grades. There is also the same use of 'role-levelling' in Trinity as is used for less proficient candidates in the Five Star test. The breadth of topic employed by the Five Star test means that there is more coverage of discrete topics. For the Trinity grades, even those at the advanced stage, switches of topic are much more restricted, with the examiner attempting to gradually build up a pseudo-authentic conversation.

The summarising tasks in the Five Star test that are prompted by listening texts are similar to the listening tasks used in Trinity Grades 10 to 12. In the case of the Trinity Advanced Levels Grades, candidates are required to listen to a passage read aloud by the examiner and then to precis what has been said. In the same way as is the case in the Five Star test, follow up discussion and extrapolation centres on the topic of the text.

Language generated

Generally speaking the kind of language generated in the two tests is similar at the outset with conversation centring on biographical information. Where the Five Star differs from the Trinity grades format is in its extensive use of task prompts via the computer leading to a number of short topic led interactions rather than the tighter overall discourse typical of Trinity Grades 7 and upwards. The format of the Five Star test also means that there are switches between prompted question-response adjacency pairs via the computer and more authentic forms of communication via extended discussion prompted by the tester. This leads to changes in the nature of the interaction which is untypical of the discourse generated by the Trinity Grade Exams, which is closer to the latter type of interaction for Five Star than the former.

The emphasis on language use in the Trinity Grade Exams is towards the functional with some attention to the structural at lower levels. The language generated in the Five Star tests is more lexically inclined, especially when it is prompted via the computer task pages. Due to the fact that the Trinity exams are basically conversational in nature, this is to be expected. Candidates have to have a good appreciation of conversational norms in order to successfully negotiate the phase of the Trinity exams and are expected to take some direction in leading the conversation from relatively low levels. This direction is essential for the higher levels - Intermediate Stage and Advance Stage. For the Five Star Test, conversational and discursual proficiency would seem to be a plus but not an absolute requirement as most of the tasks set can be completed via limited question and answer responses or summarising for the reading and listening passages.

Measurement scales

As a rough guide the two sets of test would compare in terms of candidate measurement in the following way:

Five Star test levels expressed on ESU scale 1-9 used in validation exercise	Trinity College stages	Trinity College Operational Criteria
		<i>The candidate can:</i>
ESU 1 - 2 (1 Star?)	Initial Stage (Grades 1 - 3)	<ul style="list-style-type: none"> * understand simple instructions and requests * use and recognise a narrow range of language * communicate the message at a basic level with assistance * exchange appropriate greetings and courtesies
ESU 3 - 4 (2-3 Star?)	Elementary Stage (Grades 4 - 6)	<ul style="list-style-type: none"> * understand and use basic language in everyday situations * express and ask about interests * communicate general information with some assistance
ESU 4 - 5 (3-4 Star?)	Intermediate Stage (Grades 7 - 9)	<ul style="list-style-type: none"> * understand more complex speech * use language adequately in everyday situations * communicate general ideas with greater independence * express opinions
ESU 6 - 8 (4-5 Star?)	Advanced Stage (Grades 10 - 12)	<ul style="list-style-type: none"> * use language in a variety of situations * participate effectively in an extended conversation * express opinions and explain or defend them when challenged <p>(at Grade 12 all these criteria would be approaching native speaker competence)</p>

Test reliability and validity

The reliability and validity of the Trinity Grade Spoken exams is dependent on rigorous standardisation of examiners. Since there is a human factor involved, strict reliability is relative. This is not so much the case with the Five Star test where the computer program controls the progression of a candidate through the phases of the test. In this regard the factor of inter-rater reliability is diminished, though the follow-up prompts to the task pages still mean that there may be some variance of candidate experience from test to test, as is the case with Trinity's tests.

As far as validity is concerned, the Five Star Test has a broader set of assessment objectives as compared to Trinity and measures these through pseudo-authentic tasks. Trinity's tests have the remit of objectifying oral language performance as a primary goal and attempt to assess this through three or four (at least at the higher levels) patterns of conversational interaction.

Conclusion

Though distinct in a number of ways, there are a variety of features that the Five Star Test and the Trinity College Spoken Grade Exams share. In that the Trinity tests operate as a validated assessment mechanism, the Five Star Test would also appear to function well when measured against them in terms of validity and reliability.

c) Content review against IELTS test (Kontoulis 1997)

Introduction

The IELTS Test is an internationally recognised test which provides an assessment of whether candidates are able to study or train in the medium of English. It is widely used as a language requirement for entry to many courses in further and higher education and is available at test centres around the world. Both tests are criterion - referenced; candidates are assessed on whether or not they are able to perform various tasks satisfactorily.

Skills Tested

The Five Star Test assesses four main language skills: Listening, Speaking, Reading and Study Skills, which are assessed during one test session which takes approximately 30 minutes. The IELTS test assesses Listening, Speaking, Reading and Writing, which are separate tests; the Speaking may be taken, at discretion of the test centre, on the same day or up to two days later. While all candidates take the same Listening and Speaking Modules, there is a choice of Reading and Writing modules; candidates may opt for the Academic Modules or the General Training Modules.

The total test time is 2 hours 45 minutes. Although Study Skills are not specified as a main focus of IELTS, ability in this area is essential for success in all papers, particularly on the Academic Modules.

Testing Techniques

There is a sharp contrast between the two tests in testing technique. The Five Star is an innovative test in that it is computerised but carried out by an interlocutor, who assesses the candidate's level in the initial stages and the computer selects tasks appropriate to the perceived standard of the individual. All the four skills are tested during the same session, with many tasks requiring all four and most requiring the three skills of Listening, Reading and Speaking. In several Speaking tasks the candidate is asked to first listen to instructions in Arabic and then to relate the content in English.

The tasks are based on the computer as the central testing device for the candidate, while the interlocutor combines a facilitating role with that of feeding the candidate's responses into the computer. Tasks may be repeated if the candidate feels the need for a further listening or viewing of a particular section. Certain tasks are clearly correct or incorrect, while others, mainly those requiring oral summary and responses, are assessed as poor, adequate, good or very good. The totalling of marks is done by the computer on completion of the test.

IELTS is divided into four different tests : Listening, Reading, Writing and Speaking. The Listening Module takes around 30 minutes, is recorded on tape and is heard only once; there are four sections containing between 38 and 42 questions.

The Academic Reading Module takes 60 minutes and contains between 38 and 42 questions. There are three reading passages with a total of 1,500 to 2,500 words. Texts and questions appear on a Question

Paper which candidates can write on but not remove from the test room. All answers must be answered on an answer sheet.

The General Training Reading Module, which is an option to the Academic Reading Module, follows the same format as the latter and differs only in content.

The Academic Writing Module takes 60 minutes and contains two tasks; the first requires the candidate to write at least 150 words, while the second requires at least 250 words. The General Training Writing Module, which is an option, follows exactly the same format as the above and differs only in content.

The Speaking Module takes between 10 and 15 minutes and consists of an oral interview of a single candidate by an examiner. There are five sections to this examination. Examiners work from a set of assessment criteria and guidelines and all interviews are recorded.

Elicitation Tasks and Interaction Patterns

Both the Five Star test and IELTS Speaking Module begin by the candidate being given an opportunity to talk briefly about himself. After this introduction the Five Star interlocutor proceeds through the test, eliciting general information, instructions, directions or clarification, depending on the aims of the specific task. Where a candidate shows sufficient ability, he is encouraged to speak at length about a general topic or one of relevance to his culture. This extended discourse will involve explanation, description or narration. Throughout the test, and in all tasks, there is interaction between interlocutor and candidate in an effort to negotiate meaning or clarify the nature of the task.

The IELTS candidate is guided from the introduction through an extended discourse with the examiner on a familiar topic which will also involve explanation, description or narration. The third section of the IELTS Speaking involves the candidate being encouraged to take the initiative and ask questions either to elicit information or solve a problem; this exercise is based on a task card containing 'information gap' type activities. In the fourth part of the test the candidate is encouraged to talk about his future plans and proposed course of study. This section involves speculation and expression of attitudes.

Both examinations conclude in a similar manner except that, due to the fact that the IELTS is made up of four modules, examiners are not allowed to comment on the success or failure of the test. In contrast, the Five Star interlocutor may bring the test to a close by discussing the result of the test with the candidate.

Language generated

Listening In the Five Star test the candidate is listening throughout, either to the interlocutor who explains all tasks except those spoken in Arabic, clarifies meaning and generally encourages the candidate, or to the tasks themselves. The majority of the tasks require listening usually together with other skills.

The topics are concerned with reports, news items, instructions and warnings and are monologues. A variety of questions are used, chosen from the following types: short answer questions; notes/ summary/ diagram/ flow chart/ table completion; labelling a diagram which has numbered parts matching.

IELTS Listening is a separate Module divided into four sections, the first two being concerned with social needs and the final two with situations related more closely to educational or training contexts. The sections take the form of dialogues and monologues and are all topics of general interest. The variety of questions used include those mentioned above for the Five Star test in addition to multiple choice and sentence completion.

Reading Texts used for reading in the Five Star test are quite short and based around local or international geographical or economic themes, taken from newspapers, notices, timetables or instruction manuals. A variety of questions are used, chosen from the following types: multiple choice; short-answer questions; sentence completion; matching lists; matching phrases; notes/summary/diagram/ flow chart/table completion; and classification

IELTS Academic Reading Module consists of three passages with a total of 1,500- 2,500 words. Texts

are taken from magazines, journals, books, and newspapers. They deal with issues which are appropriate and accessible to candidates entering postgraduate or undergraduate courses. At least one text contains detailed logical argument. One text may contain non-verbal materials such as diagrams, graphs or illustrations. If texts contain technical terms then a simple glossary is provided.

Texts and tasks become increasingly difficult through the paper. A variety of questions are used, chosen from all the types mentioned above for the Five Star in addition to :multiple choice; choosing from a " heading bank" for identified paragraphs/sections of the text; identification of writer's views/attitudes/claims - yes, no or not given. All answers are entered on an Answer Sheet.

The IELTS General Training Reading Module is of the same length and duration as the Academic Reading. It differs from the latter in the sources of texts; these are taken from notices, advertisements , official documents, booklets, newspapers, instruction manuals, leaflets, timetables, books and magazines.

The first section , *social survival*, contains texts relevant to basic linguistic survival in English with tasks mainly concerned with retrieving and providing general information. *Training Survival*, the second section, involves coping with a text of more complex language with some precise or elaborated expression. The third section, *general reading* , involves reading more extended prose with more complex prose with a more complex structure but with emphasis on descriptive and instructive rather than argumentative texts.

Speaking During the Five Star test the candidate is in dialogue with the interlocutor clarifying instructions, negotiating meaning and discussion. In addition, certain tasks are focussed specifically or mainly on speaking. Such tasks take the form of description , explanation, problem solving or providing short answers.

IELTS Speaking Module requires the candidate to speak at length about a general topic, elicit information/solve a problem and speculate on future situations. Materials may include a timetable, role cards explaining task, pictures/ diagrams or charts.

Writing The Five Star test does not currently assess writing. IELTS Academic Writing Module has two tasks. Task 1 requires the candidate to look at a diagram , table, or short piece of text and to present information in their own words. In Task 2 candidates are presented with a point of view, argument or problem. Topics are of general interest.

Main Differences Between the Two Tests

The computer- based Five Star test is short, concise and tests four skills within one 30 minute test session. It requires no test centre or rigid formal examination conditions and is extremely adaptable in use. In direct contrast IELTS is a formal examination which must be carried out in formal examination conditions and within approved test centres. It is a complex test lasting 2 hours 45 minutes and divided into four parts.

Five Star test has at present no written component which may be seen as a weak point. As a result of this, it would be unlikely to satisfy academic bodies who would possibly view writing as an essential skill. The outcomes of the Five Star are largely dependent on the personality and skill of the facilitator. As a result of this, IELTS, with its complex four part test, could be seen as more objective and perhaps more reliable if candidates require a general proof of language ability, particularly for academic purposes.

However, the Five Star is an extremely convenient test for the purpose for which it was developed and could be adapted for corporate use on a more international level. A further advantage is that the test result is calculated and known immediately the test is completed, while IELTS candidates must wait between two weeks and a month for their result.

Appendix V A framework for training & licensing Five Star assessors












(Parker 1997)

1. Three distinct phases in the process of licensing a Five Star assessor can be identified:

- a) Pre-training
- b) Training
- c) Post-training

a) Pre-training

Probably the most important aspect of the pre-training stage is the establishment of criteria to govern the recruitment of potential and suitable Five Star assessors. The following criteria are recommended:

FEATURE	essential	very desirable	desirable	helpful
Certificate (TESOL)				
Diploma (TESOL) or above (e.g. MA etc.)				
Native speakers of English				
Experience with other testing systems				
Experience with other oral testing systems				
Experience with counselling/interviewing in an EFL setting				
Experience of training in a Saudi setting				
Knowledge of Arabic				
EFL teaching experience				
Basic computer skills				
Exposure to Saudi culture (min. 1 year)				

b) Training There are three aspects to the training process itself:

- b1) theoretical and procedural induction
- b2) practical induction
- b3) the training manual

b1) theoretical and procedural induction

This should include the following elements in approximately this order:

Discussion of the trainee's previous testing experience, eliciting ideal features of tests and introduction to Five Star's key features.

Guided reading of key literature (eg John Pollard's articles in Testing SIG Newsletter and language Testing Update)

Familiarity with the manual

Understanding of concept of EFL proficiency and constructs (again, some published articles may be pertinent here)

Understanding of test philosophy - i.e. designed to bring the best out of the student

Understanding of the role and ideal qualities of the assessor - i.e. facilitative, sympathetic, non-threatening, supportive and reassuring

Understanding of linguistic responsibilities of assessors - i.e. avoidance of "teacher-like" language, avoidance of oversimplified forms, avoidance of language beyond the level of competence of current candidate etc.

Technical familiarisation. (Switches & buttons etc.)

Task familiarisation. (What to do/how to assess)

Guidelines for giving feedback to candidates (whether/when/how/how much) and how to avoid doing so when appropriate

Procedures for administration of test sessions (guidelines for security, comfort, procedures for managing test sessions from making appointments to databasing the results etc.)

Trouble shooting. (Contingencies for paper jams, power failures; use of the help-line etc.)

b2) Practical induction

Guided observation of edited video samples - completion of pre-set tasks (e.g. critique of assessor's/students' body language, assessor's manner, nature of balance of power etc. etc.)
 Observation of representative sample of live tests + later feedback with the assessor.
 Peer-testing + feedback
 Review of video samples and further observations of live tests if necessary (n.b. this necessity may be felt by either trainer or trainee (or, of course, both)
 Conduct of first test (observed and videoed) + Feedback. (Videoing of this test should be seen as an option rather than seeming obligatory.)
 Conduct of three* further tests (observed) + feedback
 Conduct of two* tests (observed and videoed) + feedback
 Conduct of five* tests (unobserved and videoed) + feedback (Trainee to control review of videoed tests & selection during playback)

* These numbers represent minima for licensing purposes. In the event, more may be needed in order, for example, to achieve a representative sample of levels, personality types or to satisfy the learning style of a particular trainee etc.

b3) the training manual This manual should include the following sections:

Introduction detailing the history of the Five Star test and its development
 Commentaries on guided reading*
 Tasks associated with guided reading*
 Answer keys for these tasks*
 Instructions, tasks and answer keys for video-viewing
 Statement & discussion of test philosophy and rationale
 Description of role and ideal qualities of assessors
 Notes on appropriate language control
 Guidelines on giving feedback to candidates
 Description of general technical aspects of test*
 Exercises to promote familiarity with test format*
 Description of and examples of all card types*
 Exercises to promote familiarity with all card types*
 Guidelines on the administration of tests and testing sessions
 Procedures to follow in the case of emergencies - this section might take a question & answer format
 Sample documentation - e.g. print-outs, certificates etc generated by Five Star, disclaimer forms in Arabic (photocopiable format) to facilitate videoing of tests for moderation and research etc
 Training diary (for completion by the trainee as he progresses through the stages outlined above)

*These items may in the long run take the form of separate distance learning materials

c) Post-training This area will include the following components:

Moderation. A route to achieving this would be to sample actual tests via video-recordings at a rate, say of 1 in 10 or 1 in 25 etc. or at fixed "service intervals" such as after the first ten tests and before the 15th etc. This kind of moderation would imply the existence of a chief assessor or an external examiner.
 Standardisation. This could probably best be achieved through six-monthly or annual one day re-training sessions.
 A "Court of Appeal" facility. A mechanism of sending to a chief assessor videos of actual test sessions for a second opinion. The reasons for doing this could be various such as a sense of insecurity and a desire for standardisation between meetings or the need to protect oneself against a particularly influential candidate etc
 Double marking facility. This would be very similar to the facility above but might be available to candidates in exchange for a fee???
 Five Star Newsletter. This would seem to be a sensible way to keep assessors in touch with each other and with new developments in the test.

Ray Parker
 10 January, 1997

Appendix VI Sample of Five Star test outputs: 30 candidates on 28 tasks

Cand no	4	5	6	7	8	10	11	12	13	14	15	16	17	19	22	23	24	25	26	2	28	29	30	32	33	34	36	47	Total cards	Total time
001	C					C					A				B	B	B		C	7	A								11	37.6
002	B						C		B				A	A			B		A										7	29
003	B					A	B	B	A													C	C						7	14.3
004	B					B	B		A													B	B						7	12.3
005	C					C					X																C	9	22.7	
006	B						C		A													B	C						5	26.2
007	B			A			A	A														A	B						7	10.4
008	B					A	B	B	A													B	B						7	12.8
009	B					B	B		B	B			A	B		B	A		A		A								13	55.2
010	C					B			C			A										B						A	10	20
011	C					X																					C	11	43.5	
012	B						C		A													B	C						5	11.8
013	A	C				B			B	B			A	A			B		A										9	26.7
014	B					B	B		A	B												B	B						7	29.8
015	C					X																					C	8	16.9	
016	B			B			A	B	B		A		C			A	A		B						A	B			12	24.1
017	B						C		C		A					B	B		B		C								11	17.6
018	C					C					B				B	C	C	C			C								16	47.7
019	B						C		B		A		C		B	A	B		C		A	A							15	64.7
020	B					A	B	A	B		A		B			A	A		B							A	A		12	15.3
021	B						C	C	C		B				B	C	C	C			B	A							13	34
022	B					B	B		B	B												C	B						7	17.6
023	C					X																					B	9	29.7	
024	C					C					A				A		B		C										6	18.5
025	B						C		B		A		B		B	B	B		C		A								12	33
026	B						C		B		A		B	A			A		A										8	17.5
027	C					C					C																	A	8	24.1
028	B					B	B		A	B												B	B						7	14.4
029	B					C	B				B			C	B	C	B		A		A	A							14	36.9
030	C					X																					C	8	22.6	
A scores	45	3	2	15	2	25	14	39	87	37	81	18	12	40	26	21	51	9	65	1	52	42	30	1	12	10	5	26		
B scores	157	19	1	17	1	72	72	18	72	30	59	42	33	22	53	50	68	25	32	1	13	38	44	66	6	3	4	25		
C scores	258	23	0	1	0	126	70	1	38	6	73	19	26	5	11	35	33	31	21	0	28	7	14	21	2	0	3	80		
other	0	0	0	0	0	131	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Total tests	460	45	3	33	3	354	156	58	197	73	221	79	71	67	90	106	152	65	118	2	93	87	88	88	20	13	12	131		