Reading comprehension sub-skills: teachers' perceptions of content in an EAP test

Tom Lumley

Abstract

It has been suggested that reading ability can be divided into various sub-skills, and this notion is common in ESL teaching and testing. It has, however, also been argued (Alderson & Lukmani, 1989) that teachers are unable to reach agreement about the reading sub-skills which may be tested by particular reading test items. This study begins by examining the place of sub-skills in ESL syllabus and test design, with particular attention to the enduring influence of the work of Munby (1978). The issue of teachers' perceptions of sub-skills and their difficulty, as represented in reading comprehension tests, is discussed. A framework is put forward for negotiating agreement between teachers about sub-skills tested by reading comprehension test items. Using this framework, very substantial agreement between a group of five experienced teachers of EAP was achieved in matching sub-skills to individual test items in the reading section of a test of EAP, as well as in judging the difficulty of these sub-skills. After brief discussion of the use of Rasch IRT in analysis of reading comprehension test items, the teachers' consensus regarding sub-skill difficulty level is compared to the Rasch analysis of item difficulty, and the significant correlation found gives some empirical validation to the teachers' perceptions. Implications of the findings for analysis of test content, and *for* teaching, are considered.

1. Perceptions of reading comprehension sub-skills in ESL

One way in which attempts have been made to analyse the complex process of reading is the numerous studies (e.g. Bloom, 1956; Gray, 1960; Davis, 1968; Munby, 1978) proposing the divisibility of reading comprehension into discrete sub-skills.

As a result of such studies, it is sometimes assumed by teachers and test constructors that reading can at least partially be described in

Melbourne Papers in Language Testing 1993 Volume 2.1 pp. 24-57. The Language Testing Research Centre, The University of Melbourne.

terms of 'sub-skills', or 'micro-skills'. These two terms appear to be used interchangeably in the literature, or at least without sufficient clarity of definiton for them to be clearly distinguishable.

The very concept of describable reading sub-skills is highly controversial, yet it occurs frequently in the areas of both syllabus design/materials preparation and test construction.

Reading sub-skills in syllabus design and teaching materials for ESL

Munby's (1978) framework for specifying ESP syllabus content, including its extensive list of language micro-skills, has been strongly criticised (Davies, 1981; Mead, 1982; Skehan, 1984) for various reasons, including the level of conjecture it relies upon and its lack of an empirical base, as well as its impracticality. Its impact, however, in the area of needs analysis, an important aspect of syllabus design, has been considerable.

Some writers (e.g. Grellet, 1981; Yalden, 1987) draw on Munby's work explicitly as part of the process of course design, while others (e.g. McDonough, 1984; Hutchinson & Waters, 1986) speak more generally of the requirements for needs analysis and/or for specifying micro-skills or sub-skills as part of the process of course design. The learning-centred approach to teaching which Hutchinson & Waters advocate as potentially more successful than a skills-centred approach still includes the production of a detailed language/skills syllabus.

Several points emerge from these works. The distinction between the terms 'sub-skills' and 'micro-skills' is not clear, and they often appear to be used interchangeably. Secondly, when a hierarchy or progression of skills is advocated, its shape is not defined. More generally, the issue is not that one has to be ruled by sub-skills but that they are seen as a valid component of syllabus design. An obvious point in syllabus design is that it specifies what will be taught, not necessarily what will be learned. Teachers are obliged to plan syllabuses, frequently regard (in the light of the works cited above, among many others) a needs analysis of some type as an integral part of this planning process, and consequently may wish to use this as some sort of concrete framework upon which to base their specification of course content. Inclusion of lists of sub-skills within each macro-skill is a common approach to this problem.

Sub-skills in test construction

The influence of sub-skills is also evident in language testing, as for example in the English Language Testing Service (ELTS) test (B.J. Carroll 1978; 1980) and its revised version, the International ELTS (IELTS) test (British Council/UCLES, 1989). B.J. Carroll (1980) talks of identification of language skills as part of the process of syllabus or test content specification, and of Munby's taxonomy of 54 language skills as 'one of the most fruitful sources of information for test construction' (Carroll, 1980: 32), listing eleven skills from the Munby list as suitable for testing.

Teachers are also involved in writing their own tests and in preparing students for external tests. Hughes (1989), in his handbook for teachers, discusses both 'macro-skills' and 'micro-skills'. The distinction between these two levels of sub-skills is again not made explicit, but it appears that the term 'macro-skills' here refers more to understanding the ideas in the text (information, gist, argument) while 'micro-skills' refers to recognising and interpreting the more linguistic features of the text (referents, word meanings, discourse indicators) which are the medium for the ideas, or the 'enabling' skills.

Hughes (1989: 117) sees a place for the testing of micro-skills, with the caveat that

> *'an excess of micro-skill test items should not be allowed to obscure the fact that the micro-skills are being taught not as an end in themselves but as a means of improving macro-skills'.*

The distinction is thus signalled between discrete-point and integrated approaches to language testing, a distinction related to the purpose of the test, so that in Hughes' view enabling skills may be specified in a diagnostic or progress achievement test, but not in a proficiency test.

As for identifying what it is exactly that a test tests, Hughes recommends a process of validation for tests produced by language teachers involving checking the test constructor's own perceptions of macro- and micro-skills tested with fellow teachers and with test takers.

Weir (1988) makes the same distinction between discrete-point and integrated tests in discussing the development of the Test of English for Educational Purposes (TEEP), aiming to include discrete assessment of the enabling skills in addition to integrated performance tasks. He provides a list of reading comprehension enabling skills relevant to EAP contexts (p 104).

Although the sub-skills may not be said objectively to exist, as separately identifiable, concrete elements, it is unlikely that the literature would be so full of references to them, lists of them, and continuing suggestions for their incorporation into syllabus design and test construction, if they held no inherent appeal as a working construct to many teachers involved in exactly these tasks. Of course Hutchinson & Waters are right when they state that 'learning ... is more than just a matter of presenting language items or skills and strategies' (1986: 92); but it appears that all of these elements may play a significant part in the process.

Skehan (1984) concedes that a needs analysis based upon identified sub-skills is a helpful component in syllabus design, while rejecting it as a basis for test design. He is very critical of what he terms the 'intuitive analysis' of subject areas, based upon hypothetical students, in production of the ELTS test. He concludes, however (1984: 220), with the suggestion that in the absence of empirical research in ESP contexts as a basis for test procedures

> *'It seems better to do what teachers have usually done — react to each particular set of circumstances intuitively in the manner that seems most appropriate.'*

He appears to claim, then, that the use of intuition is inappropriate for test constructors, but the only solution for teachers (although whether this is in the realms of both syllabus design and testing, or merely the former, is not made clear). This suggests that there is value in teachers' intuitions and thus lends support to the need for some sort of investigation into the kind of judgements teachers might make in identifying sub-skills in language tests.

To quote Skehan again (1984: 209), on the place of constructs in language testing:

> *'Concepts like communicative competence, sociolinguistic competence etc., are constructs. In other words, they are creations of applied linguists which, by their description and explanation, have implications and predictions for test construction. However, the vital point is that since there is a relationship between theory and practice, the performance of the actual language tests that are derived from an underlying theory becomes an important test (in the Popperian sense of falsifiability) of the credibility of the underlying theory'*

If one is to go to the trouble of specifying sub-skills, or since specifying sub-skills in language tests is clearly such widespread practice, then there should be some way of investigating whether or not they are in fact being tested in the items themselves. Skehan's point is valid when he calls for empirical studies to investigate this question.

Summary

It is clear, therefore, that the concept of reading sub-skills represents a useful working construct employed by teachers as a basis for planning syllabuses, preparing course materials and describing students' competence in language.

Furthermore, if teachers are expected a) to identify micro-skills as part of a needs analysis for their students and b) to be involved in constructing language tests, then there is a definite implication that they are capable of fulfilling these tasks to a more or less effective degree, and there is no sign that they are about to abandon these activities. What is unclear is how they do this, and what sort of reliability or validity might be attached to their judgements.

2. The relationship between reading sub-skills and test items

A variety of questions can be raised about the relationship between different reading sub-skills and between test items and sub-skills; two will be considered here:

1. Is it possible to identify a hierarchy of reading sub-skills?

2. Is it possible to relate particular reading sub-skills to individual test items?

Alderson and Lukmani (1989) investigated the questions of the existence of identifiably separate sub-skills and the idea of a hierarchy of sub-skills according to level of cognitive ability. Their study was based on items from a test used to assess the English reading ability of students at the end of their first year of undergraduate study, who had completed a course in Language & Communication Skills. The study suggested:

1) that teachers showed relatively little agreement about the sub-skills tested by a range of reading comprehension test items, leading the researchers to question the possibility of relating individual test items to identified sub-skills;

2) that the teachers disagreed considerably over the order of cognitive abilities (higher, middle or lower) demanded by the same items; and

3) that students with lower English language proficiency (as defined by performance on the test) performed as well as stronger students on items classified by the teachers (where they did agree) as requiring higher order cognitive skills, suggesting that cognitive levels were unrelated to levels of linguistic proficiency.

The last finding is unsurprising: it seems unreasonable to expect that there would necessarily be a close correlation between intelligence and cognitive development on the one hand, and level of second language proficiency on the other, which relies to a considerable extent on instruction in and familiarity with the language in question. It would seem very likely from this point of view that one should investigate students' reading proficiency in the first language and seek correlations between those measures and performance in reading tests in the second language. This point is discussed by Lee & Musumeci (1988: 184), who refer to

> *'ESL and FL learners (who) share, to a greater or lesser degree, three fundamental characteristics: 1) the cognitive maturity associated with adult age groups; 2) first language literacy; and 3) the capability of academic success.'*

Potential weaknesses are identified in Alderson & Lukmani's (1989) study in three areas: the choice of items used in the study; the content of the test; and the absence of any follow-up of the skills-item matching ratings with the teachers involved.

One claim was that the linguistically abler students performed on average no better than the weaker students on the items classified as requiring skills identified as higher order. This is surely obvious if the items show poor discriminability, as nearly half the items analysed do (6 of the 14 items examined show (p 267) discrimination values of 0.18, 0.08, 0.24, 0.1, 0.08 and 0.24). Since the establishment of adequate discriminability is a fundamental aspect of reliability of test items, these items should have been eliminated from the study.

Secondly, examination of the texts and the questions in the test suggests these low discriminability levels are not entirely surprising, as many items appear either to rely on background or cultural knowledge or to be answerable without reference to the text, suggesting they are testing things other than reading skills.

Thirdly, and most significantly for the present study, there was no exploration of why the judges made the choices they did with regard to the skills tested by test items, and no attempt was made to see where the sources of disagreement lay. In fact the point was made that the judgements were likely to have been quite different if repeated by the same teachers a week later. This highlights the need for making explicit the interpretations of the sub-skills described.

In contrast to the findings of Alderson & Lukmani (1989), a study by Brutten, Perkins & Upshur (1991), investigating whether certain ESL reading comprehension skills were shown to lag behind others, as measured by performance on the TOEFL, found a high level of agreement between four raters about the skills tested by individual test items, using the Iowa Test of Basic Skills taxonomy of reading skills (Hieronymus, Hoover & Lindquist, 1986).

3. The study

In the light of the importance of the question of reading sub-skills in ESL teaching and testing, and the conflicting research findings reported above, the issue merits further investigation.

This study firstly analyses the attempts of a group of experienced EAP teachers to match particular descriptions of micro-skills to individual reading comprehension test items. The intention is to examine what degree of consistency or consensus they are able to achieve in their judgements, and how such consensus may be reached. The suggestion is that if an appropriate methodology is adopted, then it may be possible to challenge the finding of Alderson & Lukmani (1989) that judges are unable to agree on the sub-skills tested by particular items, and hence lend support to the contrary finding of Brutten et al. (1991). The judges' agreement about relative difficulty of the sub-skills considered will also be examined.

Two main questions are considered in this part of the study:

1) Do the same teachers perceive a common hierarchy of difficulty amongst the sub-skills?

2) Is it possible for a group of experienced EAP teachers to reach agreement upon sub-skills tested by individual test items in a test of reading comprehension?

A further issue investigated later in the study concerns the possibility of validating the teachers' perceptions by comparing them with an analysis of the performance of candidates on test items using Rasch item response theory (IRT).

The Test

An EAP test was developed for the University of Melbourne (Brown & Lumley 1991a, 1991b; Lumley & Brown, 1991; Lumley, 1992), designed principally to provide diagnostic information about the language proficiency of non-English-speaking background (NESB) students at the University, from all faculties. It was possible that the test would also ultimately be used for screening purposes. The test included writing, listening and reading sub-tests. For this study, items in the reading comprehension sub-test were analysed. This sub-test included 58 items based on two texts with a total length of approximately 1500 words.

Clapham (1991) has pointed to the likelihood of the influence of academic area and background knowledge upon test scores in university language proficiency tests. An attempt was made to reduce the factor of background knowledge by choosing reading texts relating to common environmental issues, where concepts were presented, explained and discussed. The conceptual level of the texts was approximately that of the final year of secondary school, i.e., at or below the presumed conceptual level of the test candidates.

There was a variety of item types: short answer, cloze, multiple choice, matching, true/false, completing a flow-chart, labelling maps.

The subjects

The test was trialled on 3 groups of NESB subjects (N = 158):

1) (N = 90) a group of overseas students (mainly undergraduate) already at the University, from a wide range of language backgrounds;

2) (N = 50) a group of students from a local English language centre, from Chinese, Indonesian, Japanese, Korean and Thai language backgrounds, undertaking short intensive EAP courses in preparation for the IELTS test and/or tertiary study;

3) (N = 18) a group of post-graduate students from Eastern Europe, with academic backgrounds in business or economics, studying for a post-graduate qualification in business administration.

These three groups were deemed to represent a range of levels of ESL proficiency, all of whom would nevertheless require an ability to use English for academic purposes, and were assumed to share the three characteristics described by Lee & Musumeci (1988) referred to above.

Test Analysis

Test results are summarised in Table 1. Satisfactory estimates of test reliability were obtained for the test as a whole. This is important in the context of Alderson and Lukmani's study, where such an analysis is not reported. That the trial subjects represented a range of abilities was borne out by performance on the reading sub-test, with a spread of scores from 4 to 57 out of a total of 58, a mean score of 33.85, and a standard deviation of 15.24.

Mean score	33.85
Std. Dev.	15.24
Std. Error	1.212
Spread of scores	4 – 57
K-R 21	0.95
Rasch person separation index	2.90
Rasch reliability of person estimates	0.89
Rasch item separation index	4.16
Rasch reliability of item estimates	0.95
N	158

Table 1. Analysis of the reading sub-test

Rasch analysis, using the program QUEST (Adams and Khoo, 1990, 1992) showed four items as misfitting, with infit meansquare values above the acceptable limit of 1.3. Classical analysis showed that discrimination and facility values obtained were acceptable.

Procedure

In effect, the process described in this section constituted a post hoc content analysis, employing a procedure of negotiation and justification to establish common interpretations of sub-skills descriptions, similar to those used in determining criteria and interpretations of those criteria in tests of writing for the ELTS revision (Hamp-Lyons 1986) and of writing and speaking for the Occupational English Test for health professionals (McNamara 1990a, 1990b).

Existing lists of sub-skills

Initially, the work of Gray (1960), Bloom (1956), Davis (1968) and Lunzer, Waite and Dolan (1979), was examined, in an attempt to identify a suitable set of reading sub-skill descriptions for use in this study. In each case, the descriptions given appeared too general and undefined to describe adequately the items tested in the University reading test. These studies addressed the question in relation to the development of children's reading in English as mother tongue. Since it is uncertain whether the reading process is similar for adults reading in their second language, it was decided to consult studies which considered the issue from the point of view of adult learners of ESL. Munby's (1978) more detailed list of 19 reading micro-skills was examined, as was the list of skills (possibly) tested by reading comprehension questions in the exam set at the end of the Bombay University Communication Skills Course (Alderson & Lukmani 1989: 260).

This last list includes an explanation under each skill heading of what ability (or in the case of the first one, what sort of knowledge) is being assessed in each case. There is thus more of a focus on how these skills might be tested than in Munby's list, which was designed as an aid for syllabus design and teaching. Because of this focus on testing, the sub-skill descriptions used in the Alderson & Lukmani study, those defined by the Bombay Communication Skills Group (BCSG) were considered by the author as a possible model for sub-skill descriptions for items in the University test. However, it proved too difficult to distinguish clearly terms such as 'analysis', 'interpretation' and 'inference'. Looking at the tasks Alderson & Lukmani (1989: 258) quote from Adams-Smith (1981), as testing analysis, synthesis and evaluation, these seem to be testing things other than reading, and would appear to sit more comfortably in a test of writing.

An attempt to shed light on these distinctions was made by a group of eight Applied Linguists from the University of Melbourne (including the author), who considered a group of six (of the most difficult) items from the reading test and attempted to determine the sub-skill(s) from the BCSG list needed to answer each item. After a protracted period of discussion, it became apparent that there was no clearly accepted interpretation of the meaning of each sub-skill category, nor much agreeement about their allocation to individual test items. This appeared to confirm the centrality of the issue of defining the terms used in sub-skill descriptions: without a common understanding of their meaning, there appeared to be no hope of a meaningful attempt to assign sub-skills to test items.

A further possible explanation for the lack of agreement among raters was that although this was the first time they had participated in such an exercise, they were being presented with complex skills and asked to associate them with a group of the most difficult items in the test. It seemed that a more logical starting point was with the simpler items, moving towards the more complex, defining and elaborating the procedure needed to establish agreement, rather than beginning with an examination of the most difficult items.

Final selection of test items for analysis

The six items referred to above were not considered further. In addition, seven items with poor discrimination, or poorly worded, were rejected for the present study, as were all the cloze items, on the grounds that as integrative test items requiring the use of a wide range of linguistic skills simultaneously (Oller 1975; Bachman 1985; Jonz 1987, 1990) it would not be possible to separate and define the skills needed in this section. Twenty-two items from the reading test were selected, covering the full range of difficulty, (logit values -1.875 to 1.875; discrimination levels according to classical analysis in the range 0.55 to 1.0 and facility levels from 0.89 to 0.25).

Development of the present list of sub-skills

The wording of sub-skill descriptions appeared as a crucial factor if they were to be meaningful to those analysing the test. It was considered important to use descriptions including as much detail as possible, in order to provide maximum clarity. No existing taxonomy proving appropriate, the author decided to generate a new set of descriptions relevant to this set of items. In order to avoid making unrealistic claims about what items appeared actually to be testing, some of the descriptions produced related as much to the task as to a supposed underlying linguistic skill.

The list of sub-skills eventually developed by the author to describe the 22 items under consideration is reproduced in Appendix A.

The raters

One of the criticisms of Munby's list of micro-skills was that it was merely the product of his own speculation (his own attempt at guessing about psycholinguistic processes), and never subjected to consideration by other applied linguists. For this study a group of five raters was assembled, to consider the test items and the sub-skill descriptions produced by the author. All raters were qualified ESL/EFL teachers with at least five years' experience, much of it with students preparing for or engaged in tertiary study; all had fully or partially completed a Master's degree in Applied Linguistics; all were involved in language test construction; the group included the two test developers. All members of the group had completed the reading test before the rating session.

The rating session

The procedure during the session was as follows:

1) The group rated the list of sub-skills in terms of perceived difficulty on a four-point scale (A–D, with A representing the easiest).

2) They rated the selected items on the same scale.

3) Group members then selected, from the list of sub-skills supplied, what they considered to be the single, highest level skill required to answer the question. They were to interpret this as the skill, in their judgement, without which it would not be possible to answer the question: although answering a question might require the operation of more than one sub-skill, group members were encouraged where possible to limit their choice to the single most important skill.

4) Group members allocated a sub-skill (or sometimes two) to the first item from the 22 items under consideration. Judgements were compared and justified to the group. After this debate, which involved focussing on people's perceptions of the likely reading process, how the answer might have been arrived at, and on clarifying the meaning of the wording of the sub-skills chosen, each person was again asked to allocate a sub-skill to the item. This process was repeated for three more items, of varying levels of difficulty, in order to establish a level of agreement about firstly the procedure and secondly the focus and interpretation of the sub-skill descriptions.

5) Sub-skills were then matched to the remaining items by each group member, with the same process repeated at the end of the session with each item.

A number of major points emerged from this exercise, including:

1) The importance of determining the focus of each sub-skill. For example, the difference between

4. Explaining a fact with:

4.1. a single cause

4.2. multiple causes

and

6. Analysis of the elements within a process, to examine methodically their causal / sequential relationship.

was initially unclear. The group defined the focus of no. 4 as a single fact resulting from one or more causes, and the focus of no. 6 as a process, where there was a relationship between the stages in this process.

2) It was recognised that knowledge of vocabulary, referred to in sub-skill 1:

Dealing with relatively uncommon vocabulary: matching of words/phrases referred to in text with given equivalent meanings,

was fundamental to answering all items, but impossible to describe or measure: it was important to consider the concept of one or more particular sub-skills being necessary but not sufficient for answering the question, and the necessity therefore of concentrating on what actually led one to the answer.

3) Some sub-skills would definitely occur at several or all levels: the most noticeable example of a sub-skill judged in this way was sub-skill 5:

Selecting a phrase as summarising the main topic of a text..

4) Two sub-skills commonly listed and taught as important in reading comprehension, skimming and scanning, were needed repeatedly throughout the test, but could not be identified as central to particular items. This is consistent with the findings of Lee & Musumeci (1988: 175) in their analysis of the ACTFL Proficiency Guidelines, that skimming and scanning could not be *isolated* as reading skills, but that in completing reading tasks involving these strategies, readers would be employing a further reading skill.

5) It was necessary to alter slightly the wording of some sub-skills and to add a ninth sub-skill to the existing list (for which, therefore, no level of perceived difficulty was identified).

6) Potential confusion remained between sub-skill 9:

understanding grammatical and semantic reference and sub-skill 3:

> *identification of information in the text, clearly stated but in paraphrase (or where no key word occurring in both text and question will lead directly to the answer).*

In sub-skill 9, the distinction between grammatical and semantic reference is hard to draw, for which reason they were put together. However, since the result of understanding these kinds of references is to enable the reader to identify relevant information, as described in sub-skill 3, possibly sub-skill 9 should be seen as part of sub-skill 3.

4. Results

Using a scale from A (easiest) to D (most difficult), the raters demonstrated varying levels of agreement in relating the sub-skills to each other in terms of inherent difficulty. Intermediate points were also used, so that A+ indicates a sub-skill perceived as harder than A but easier than B. No guidelines were given to the group as to how they should interpret the levels A to D; the decision to mark intermediate positions was made independently by the raters, seen by them as necessary to reflect their perceptions of the skills listed.

Skill	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Consensus
1	A	A	A	B	B	A–B
2.1	A	A	A	A	A	**A**
2.2	A+	A+	A	A+	A+	**A+ (A)**
2.3	B	A++	A	A+	A+	A–B
3.1	B	B	B	B	A+	**B (A+)**
3.2	B+	B+	C	B+	B+	**B+ (C)**
3.3	B+	B+	B+	C	B+	**B+ (C)**
4.1	C	B	B	C	B	B– C
4.2	C+	B	D	C+	C	none
5	C–D	B–D	C	A–D	C–D	**varies**
6.1	D	B	B	C+	C	none
6.2	D+	C	D	D+	D	C– D+
7	D	D	D	D	D	**D**
8	C	C	B	C	C	**C (B)**

Table 2. Difficulty of subskills

Notes for Table 2:

Figures in bold represent cases where 80% or greater agreement between judges was produced, with other ratings shown in parentheses; figures in light print in the 'Consensus' column represent the range of ratings given, although if there was a difference of more than one level, there was considered to be no consensus.

Table 2 shows that for two sub-skills there was 100% agreement about the level of difficulty; for five there was 80% agreement, with the fifth rating no more than one band higher or lower; for sub-skills 1, 2.3 and 4.1 opinion was split between two adjacent band levels; and for sub-skill 5 there was an 80% consensus that the skill would occur at a range of levels of difficulty, although there was disagreement at how easy such a skill might be. Thus for 11 of the 14 sub-skills described there is seen to be substantial agreement about inherent levels of difficulty. It is only sub-skills 4.2 and 6.1 that reveal strong disagreement, and to a lesser extent sub-skill 6.2.

To test the significance of the concordance between teachers' perceptions of the sub-skills' difficulties, each judge's ratings were ranked (Kendall's W = 0.849; Chi-Squared = 50.934, $p < .001$). No rating was possible for sub-skill 5, and it was omitted from this analysis.

Item	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Consensus
1	5	5	5	5	5	5
2	5	5	5	5	5	5
5	(6.2&6.1)	(2.2)	(1/3.2/6.1)	(3.1)	(3.3)	
	4.1	4.1	4.1	4.1	4.1	4.1
6	2.1	2.1	2.2	2.1	2.2	2
7	2.1	2.1	2.1	2.1	2.1	2.1
		& 3.2				
8	None	(3.3 &1)	(2.3 &1)	(2.3 &1)	(1)	
	9+1	9+1	9+1	9+1	9+1	9+1
9		(2.3/3.3?)				
	4.2	4.2	4.2	4.2	4.2	4.2
10	2.1	2.1	2.1	2.1	2.1	2.1
14		(2.1/3.1?)				
	3.1&3.3	3.1	3.1	3.1	3.1	3.1
19–22	6.2	6.2	6.2 & 8	6.2	6.2	6.2
23	(1&7)		(1&7)	2.1?		
	7 & 1	7	7&1 or 2.1	7	7	7
31	2.1	2.1 &??	2.1	2.1	2.1	2.1
32	2.1	2.1	2.1	2.1	2.1	2.1
33	3.1	3.1& 3.3?	3.1	3.1	3.1	3.1
34	3.1	3.1	3.1	3.1	3.1	3.1
35	8	8	8	8	8	8
36	8	8	8	8	8	8
37	8	8 & 3.3	8	8 & 2.1	8	8
38	7 & 5	7	7	7	7	7

Table 3 Reading sub-skills matched to each item

Notes for Table 3:

1) The figures in **bold** represent the consensus reached by the five raters as to the micro-skill tested;

2) the figures in parentheses represent micro-skills initially selected by raters;

3) '&' indicates that additional micro-skills were selected as nec. by raters.

4) '?' indicates that the rater was uncertain about whether a sub-skill was needed, or which one.

5) 'None' indicates that the rater was unable to find a sub-skill in the list provided which adequately described what a test taker might do.

The results of the process of assigning sub-skills to items are shown in Table 3. This table shows substantial change in raters' selection of the principal sub-skill required to answer each item, as a result of the group discussion. This is particularly noticeable in the cases of items 8, 5 and 23. For item 6, although the general skill required was seen as sub-skill 2, 'identification of information explicitly stated', agreement was not reached as to whether this was sub-skill 2.1 or 2.2, in other words, on the degree of complexity of the relevant sentence. As can also be seen, in some cases one or more raters wanted to add a second skill to those listed, as also essential.

The principal finding here is, however, that the five raters showed that, as a result of the discussion of items and clarification of meaning of sub-skill descriptions, they *were* able to reach almost complete agreement on which was the principal skill necessary to answer each of the 22 items considered.

5. Rasch IRT analysis and reading sub-skills

In this section the use of IRT in analysis of sub-skills in tests of reading comprehension is discussed. The possibility of IRT analysis providing empirical validation of the teachers' perceptions is considered.

Use of IRT in tests of reading comprehension

IRT has been used in analysis of first language tests of reading comprehension which claim to test particular reading sub-skills. These tests include the National Assessment of Educational Progess (NAEP) and the Iowa Basic Skills Test in the USA, and the Tests of Reading Comprehension (TORCH) in Australia. All three tests claim to test particular sub-skills.

In the case of the TORCH (Mossenson et al., 1987) 302 items from 14 reading tests were arranged along a continuum of increasing difficulty. The items were grouped in order to produce eleven distinct kinds of tasks or sub-skills, although it is unclear what process was used to relate the items to the tasks. It should be possible to use this scale, when analysing student performance, to provide diagnostic information for use in teaching. These tests are tests of L1, used (mainly) for children in primary schools. This may represent a different case from that of an EAP test of second language written for adults, because the issue of the influence of conceptual development upon performance in native-speaking children is significantly replaced by other issues such as cultural and background (including educational) knowledge in the adult NESB students.

In the development of one test of ESL, the Interview Test of English as a Second Language (Adams et al., 1987; Griffin et al., 1988) a set of linguistic (basically grammatical) objectives was produced, which when subjected to item response analysis led to the development of a set of test items which were considered to represent 'the continuum underlying the development of language skills' (Griffin & Nix, 1991: 79). If used as a set, it is claimed that these items can be used for the purposes of evaluation of language programmes or diagnosis of students' strengths and weaknesses, and that

> *'This approach can be applied to any area of learning for which it is possible to analyse the stages of development, to define the evidence of those stages of development, to define the means of observing that evidence and to scale the data to define the trait'* *(p 79)*

The test has received a mixed response (eg, Spolsky, 1988; Hamp-Lyons, 1989; McNamara, 1990a), but one basic problem may be the complexity of validating the stages described in this approach.

Rasch analysis in language testing is unique in mapping student ability and item difficulty on the same scale, thus enabling maps of student performance to be produced, from which ability statements may be developed, based on sub-skill descriptions (Brown et al., 1992). It is possible that such sub-skill descriptions could be produced by teachers, following a procedure similar to that described earlier in this study. Resulting analyses of strengths and weaknesses of both individuals and groups have the potential to provide useful guidance for teachers in planning their teaching. Such an approach may be appropriate in the context of the TORCH scale of reading tasks, as described above, representing increasing reading ability.

IRT analysis as validation of teacher perceptions

For such a scale to be usable it is necessary to consider whether the sub-skills identified by teachers as related to particular test items do in fact appear to fall into groups, so they may be placed onto a scale of increasing difficulty.

The question now to be investigated is: do items identified by the group of teachers as requiring the same sub-skills occur at roughly the same level of difficulty as each other, according to the Rasch analysis?

Using the data in Table 4, the correlation was calculated between the difficulty of each item, as shown by the logit value, and the difficulty of each sub-skill, according to the ratings given by the group of teachers, producing the following result:

$r = 0.716$ r-squared $= 0.513$

$df = 13$ $p < 0.01$ (two-tailed)

Sub-skill	Item no.	logit value (difficulty)	Skill level consensus	Skill rating
2.1	7	-1.75	A	1
2.1	10	-1.875	A	1
2.1/2.2	6	-1.5	A/A+	1
2.1	31	-0.625	A	1
2.1	32	-0.125	A	1
3.1	14	-0.625	B	2
3.1	33	1.0	B	2
3.1	34	1.375	B	2
4.1	5	-1.5	B–C	⊗
4.2	9	0.5	no agreement	⊗
5	1	1.875	A–D	4
5	2	0.5	A–D	4
6.2	19	-0.5	C–D+	⊗
6.2	20	-0.875	C–D+	⊗
6.2	21	-0.875	C–D+	⊗
6.2	22	-1.0	C–D+	⊗
7	23	1.375	D	4
7	37	0.875	D	4
7	38	0.25	D	4
8	35	0	C	3
8	36	-0.125	C	3

Table 4 . Sub-skills: IRT values and teachers' consensus

Notes for Table 4:

1) the sub-skills under consideration are listed in column 1;

2) the items identified by the group of teachers as testing each sub-skill are listed in column 2;

3) column 3 shows the logit value produced by Rasch analysis of the test;

4) the perceived difficulty of the sub-skill (where A is seen as the easiest sub-skill), according to the group of teachers, is shown in column 4;

5) the teachers' ratings have been re-categorised, on a 4-point scale, in column 5: where full agreement was not expressed about sub-skill difficulty the sub-skill was excluded from the calculation.

6) sub-skill 5 ('Selecting a phrase as summarising the main topic of a text') was the only one specifically identified as dependent in terms of difficulty upon the complexity of the text, varying in potential difficulty from A to D, though generally towards the difficult end of the continuum. In this case, since the items tested by it were rated universally by the group of teachers as amongst the most difficult, it was seen as appropriate to assign it the highest value on this scale.

This significant correlation gives some empirical support to the validity of the teachers' perceptions, in that they would appear to have in some way identified elements common to groups of items, or at least part of what makes one item more difficult than another.

The data in Table 4 can be represented as in Figure 1.

Here the items at each skill level (1, 2, 3 or 4) can be seen to fall into broad, though overlapping, bands that demonstrate a general tendency of increasing difficulty, as represented by the logit values of the items.

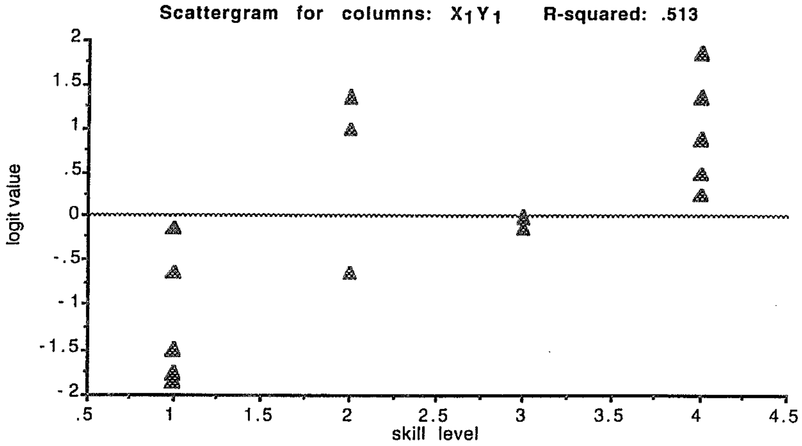

Figure 1. Representation of Table 4.

6. Summary of findings

1) If the sort of procedure described in the earlier part of the present study is followed, it seems possible for a group of teachers to reach a high level of agreement about sub-skills tested by particular test items. This group also share common perceptions about the relative difficulty level of the sub-skills described in this study.

2) There is a significant correlation between the teachers' perceptions of the difficulty of each sub-skill and the logit values obtained from the IRT analysis for items identified as testing the same skills. Some empirical justification is thereby given for the teachers' perceptions of the content of this test. Furthermore, the sub-skills considered in this study, from this reading test, are seen to fit into broad bands of increasing difficulty.

7. Discussion

These findings have implications for the concept of reading sub-skill classification, as well as for the role of sub-skills in the interpretation of test performance and in teaching. Issues affecting these questions include the reading process, the test taker's background knowledge and cognitive style, the wording of sub-skill descriptions, the variation between texts, and the purpose of the test.

With reference to these findings, it must first be said that obviously teachers cannot judge what happens inside students' heads. One difficulty with the sort of study described here is the impossibility of knowing exactly what students do when they read, and whether it differs substantially from the teachers' perceptions of what they do, or from one reader to another. There are issues here that can to some extent be addressed in introspective studies where students describe the reading or test-taking process (see for example Hosenfeld, 1984; Harri-Augstein and Thomas, 1984; Cohen 1988). To take one very simple example, in deciding how much a reader might read in order to feel reasonably confident of having answered a question correctly, one can only speculate that individual differences in such things as cultural background, subject knowledge, attitude to the test and cognitive style might lead some readers to stop as soon as an apparently correct answer is obtained, while others might want to read on, an indeterminable amount, to confirm their answer.

A further example regarding the question of background knowledge in the context of some of the sub-skill descriptions considered in this study is that different readers are likely to see different levels of explicitness in a text, or identify different 'keywords'. This is a matter about which only a study of the test-takers could give information. If reading is at least partially a top-down process, then any text will be to some extent the reader's re-interpretation of his/her own experience, which will vary tremendously from one individual to another.

There is also the point that what is inference for one person is simply knowledge or recognition for another (Tuinman 1979). This would seem to apply particularly in the realm of vocabulary (and consequently in those sub-skills concerned with explicitly stated information or paraphrase), and would partly explain why it was impossible to formulate a satisfactory sub-skill describing vocabulary knowledge or inference of meaning.

Despite these cautionary notes, it *has* been shown that teachers are able to agree in their speculation about the possible processes involved in answering test items. This speculation may be informed by some intuition about what makes an item hard, or what differentiates it from other items. In some cases this may relate to the task set in the test rather than the skill a reader would use in everyday life; or perhaps the task and skill together. It is important to recognise that the sub-skills identified by the teachers are not claimed to be the only things tested by the test items, but rather an expression of what they considered to be skills without which the question could not be answered.

Whatever it is that is being described may as yet be unclear, but the fact that consistency of judgement can be shown to take place gives some empirical support for the procedure of mapping reading skills from test content as used in, for example, the TORCH.

Where disagreement does occur between judges about the difficulty of a sub-skill, or about which sub-skill is tested by an item, given the apparent centrality of exact descriptions of sub-skills, and of defining the terms used within them, this may be due to inadequacies in the phrasing of the sub-skill. Therefore the descriptions given for sub-skills 4.1, 4.2 or 6.2, for example, may not be helpful in describing any of the items in the test examined here. There is potential for much more extensive investigation of this question, including of course the major issue of the development of a theoretical basis for producing sub-skill descriptions.

With regard to the diagnostic value of a sub-skills analysis of test performance, then without making unreasonable claims for what a reader can or cannot do, the information yielded by the identification of any sub-skill as inadequately developed in a group of students could perhaps signal to a teacher a useful area of work as a focus for teaching. An extension of this idea, using individual maps produced from analysis of individual student performance, in order to provide more detailed diagnostic information for teachers, seems also to have potential value. There remain the questions as to whether this implied use for diagnostic tests could be generalised to proficiency tests, in the light of Hughes' (1986) and Skehan's (1984) cautions, and whether these two types of test are necessarily entirely distinct.

The claim may of course be made that there is no such thing as skills that are inherently easier or more difficult, but that the difficulty of any sub-skill will be completely dependent upon background knowledge and the properties of the text used in the test. This raises a number of issues, including those relating to individual learner differences referred to above.

One factor affecting the difficulty of any question, and equally of the sub-skill(s) related to it, is the readability/complexity of the text. The fact that teachers did agree about level of difficulty for the sub-skills suggests that they may have been making decisions based only upon the test studied here, or else considering or assuming that the sub-skills would be applied to texts at the same level as each other. The question of readability is a common one in test design and evaluation, a point made strongly by J.B. Carroll (1986) in his discussion of the NAEP, where he called for objective data on the reading difficulty or readability of the texts used in the test, since ability to use the so-called 'enabling skills' would depend on such things as the readability, vocabulary level and syntactic complexity of the texts. It is unclear how these might be characterised. The question of how far sub-skills can usefully be bound to individual contexts assumes central importance, as signalled by Skehan (1984: 216), in discussing the possibility of criterion-referenced testing:

> *'The problem is, fundamentally, that any language performance that is worthy of interest will be complex and multidimensional. Because of this it will be impossible to state what the criterion (for assessing any language performance) is for any except a small number of tightly-defined contexts'*

A partial attempt is made to address this issue in the test examined in this study by the specification that the test should be composed of academic reading texts similar to those commonly encountered in the final year or two of high school, which may at least give some indication of the conceptual level assumed, even though the kind of texts encountered by students at these levels may vary widely.

As for the findings in the latter section of the study, a number of reservations need to be put forward.

The present study offers data on only a small number of items. Further investigation is needed of the efficacy of the procedure put forward here, using a much larger sample of items and a wider range of sub-skill descriptions.

The picture that emerges in Diagram 1 above of bands with a considerable degree of overlap is not in the least unexpected: it would be surprising if one were to claim to show that one skill had to be fully acquired before the next could be mastered. More likely is the position suggested by Griffin & Nix (1991) of gradually emerging mastery of linguistic skills of increasing difficulty, as ability increases. What is unclear from this picture is how widely the bands may extend: this, too, would require further investigation. Further research might begin to establish how far changes in text and context will affect the estimated level of difficulty of the sub-skills described, and under what circumstances particular micro-skills cluster together.

Linked with this is the issue of the influence of test method facet (Bachman 1990): to what extent do the item type and formulation of the question affect reader performance, and to what extent is their performance determined by the text itself? There is much scope for further research in this area, employing a variety of testing methods with the same texts, or parallel methods with different types and levels of texts. Lee & Musumeci (1988) refer to this question in relation to the texts and skill descriptions used in the ACTFL Proficiency Guidelines as well as in their own study of the reading performance of learners of Italian.

One obvious point emerging from this study is that a group of people aiming to achieve consensus are likely to do so. Another approach to the question therefore would be to seek the sort of agreement achieved in this study in a session similar to the one described; the same raters would then analyse a further set of items, with a text seen as comparable, or with additional items relating to the same text, to see whether they collectively assigned the same sub-skills to these new items.

8. Conclusion

The continuing importance of reading sub-skills in syllabus design, teaching materials and test construction has been considered in this study. It is not suggested that these sub-skills 'exist' in any tangible way, but rather that they represent a useful construct with which teachers and test constructors may work.

An investigation was carried out of the level of agreement shown by a group of five experienced ESL teachers on 1) the identification of selected reading sub-skills with particular items in a test of reading comprehension, and 2) the relationship between these sub-skills in terms of perceived difficulty. In the case of the first question, almost complete consensus was reached between the raters for all items, following a procedure involving discussion and definition of terms. With regard to the second issue, a high level of concordance was demonstrated between their perceptions.

An analysis of the relationship between the teachers' perceptions of sub-skills tested and the logit values produced by Rasch analysis of the items identified as testing these sub-skills, showed a significant correlation. It appears possible from the (admittedly limited) data in this study to perceive bands of increasing difficulty associated with a number of the particular sub-skills examined.

These findings collectively lend some empirical support to the value of using teachers' judgements in examining test content, and to the procedure followed in some areas of test development (eg Mossenson et al 1987), involving mapping skills from test content. The judgements they make about linguistic matters in test design and content validity also have significance for teaching.

9. References

Adams, R.J. & S.T. Khoo (1990) *TITAN: the interactive test analysis system*. Hawthorn, Victoria: Australian Council for Educational Research.

Adams, R.J. & S.T. Khoo (1992) *QUEST*. Hawthorn, Victoria: Australian Council for Educational Research.

Adams, R.J., P.E. Griffin, & L. Martin (1987) 'A latent trait method for measuring a dimension in second language proficiency.' *Language Testing*, 4, 9–27.

Adams-Smith, D.E. (1981) 'Levels of questioning: teaching creative thinking in ESP.' *English Teaching Forum*, January 1981, 15–17, 21.

Alderson, J.C. & Y. Lukmani (1989) 'Cognition and reading: Cognitive levels as embodied in test questions.' *Reading in a Foreign Language*, 5, 2, 253–270.

Alderson, J.C. & A.H. Urquhart (1984) *Reading in a Foreign Language*. London: Longman.

Bachman, L.F. (1985) 'Performance on cloze tests with fixed-ratio and rational deletions.' *TESOL Quarterly*, 19, 535–556.

Bachman, L.F. (1990) *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bloom, B.S (1956) *Taxonomy of Educational Objectives. Book 1: Cognitive Domain.* London: Longman Group Ltd.

British Council / UCLES (1989) *Specifications for IELTS.* Cambridge: British Council/UCLES.

Brown, A. & T. Lumley (1991a) *The University of Melbourne ESL Test: final report.* Language Testing Centre, University of Melbourne.

Brown, A. & T. Lumley (1991b) Some practical applications of Item Response Theory in language programmes. Paper presented at 4th ELICOS Association Conference, Monash University, August 21–23.

Brown A., C. Elder, T. Lumley, T.F. McNamara & J. McQueen (1992) Mapping abilities and skill levels using Rasch techniques. *Melbourne Papers in Language Testing*, 1, 1.

Brutten, S.R., K. Perkins & J.A. Upshur (1991) 'Measuring growth in ESL reading.' Paper presented at the Thirteenth Annual Language Testing Research Colloquium, Princeton, New Jersey, March.

Carroll, B.J. (1978) *An English Language Testing Service: Specifications.* London: British Council.

Carroll, B.J. (1980) *Testing Communicative Competence: An Interim Study*. Oxford: Pergamon Press.

Carroll, J.B. (1986) 'Scales and other problems in the NAEP reading assessment: critical comments.' A paper commissioned by the Study Group on the National Assessment of Student Achievement and cited in Appendix B to their final report *'The Nation's Report Card'*. Chapel Hill: University of North Carolina.

Clapham, C. (1991) 'The effect of academic discipline on reading test performance.' Paper presented at the Thirteenth Annual Testing Research Colloquium, Princeton, New Jersey, March.

Cohen, A.D. (1988) The use of verbal report data for a better understanding of test taking processes. *Australian Review of Applied Linguistics,* 11: 30–42.

Davies, A. (1981) 'Review of Munby, J. (1978) *Communicative Syllabus Design.' TESOL Quarterly,* 15, 3, 332–6.

Davis, F.B. (1968) 'Research in comprehension in reading.' *Reading Research Quarterly,* 3, 499–545.

Gray, W.S. (1960) 'The major aspects of reading' in Robinson, H. (ed) *Sequential Development of Reading Abilities,* Supplementary Educational Monographs No. 90. Chicago: University of Chicago Press, 8–24.

Grellet, F. (1981) *Developing Reading Skills: A Practical Guide to Reading Comprehension Exercises.* Cambridge: Cambridge University Press.

Griffin, P.E., R.J. Adams, L. Martin & B. Tomlinson (1988) 'An algorithmic approach to prescriptive assessment.' *Language Testing,* 5,1, 1–18.

Griffin, P.E. & P. Nix (1991) *Educational Assessment and Reporting: A New Approach.* Sydney: Harcourt Brace Jovanovich.

Hamp-Lyons, L. (1986) 'Testing writing across the curriculum.' *Papers in Applied Linguistics — Michigan,* 2, 1, 16–26.

Hamp-Lyons, L. (1989) 'Applying the partial credit model of Rasch analysis: language testing and accountability.' *Language Testing,* 6,1, 109–118.

Harri-Augstein, S. & L.F. Thomas (1984) 'Conversational investigations of reading: the self-organised learner and the text' in Alderson & Urquhart (1984).

Hieronymus, A.N., H.D. Hoover & E.F. Lindquist (1986) *Preliminary Teacher's Guide. Multilevel Battery Levels 9–14.* Chicago: Riverside Publishing Company.

Hosenfeld, C. (1984) 'Case studies of ninth grade readers' in Alderson & Urquhart (1984).

Hughes, A. (1986) 'A pragmatic approach to criterion-referenced foreign language testing' in Portal (ed.) (1986)*Innovations in Language Testing*. Windsor: NFER-NELSON, 31–40.

Hughes, A. (1989) *Testing for Language Teachers.* Cambridge: Cambridge University Press.

Hutchinson, T. & A. Waters (1986) *English for Specific Purposes: A Learning-centred Approach.* Cambridge: Cambridge University Press.

Jonz, J. (1987) 'Textual cohesion and second-language comprehension.' *Language Learning*, 37, 3, 409–38.

Jonz, J. (1990) 'Another turn in the conversation: what does cloze measure?' *TESOL Quarterly*, 24, 61–83.

Lee, J.F. & D. Musumeci (1988) 'On hierarchies of reading skills and text types.' *The Modern Language Journal*, 72, 2, 173–187.

Lumley, T. (1992) Reading comprehension sub-skills in an EAP test: teachers' perceptions of what is tested. Unpublished MA thesis, University of Melbourne.

Lumley, T. & A. Brown (1991) The development of a diagnostic EAP test: the application of Item Response Theory. Paper presented at the 16th Applied Linguistics Association of Australia Conference, James Cook University, Townsville, September 30–October 2.

4. Explaining a fact with:

    4.1. a single cause

    4.2. multiple causes

5. Selecting a phrase as summarising the main topic of a text.

6. Analysis of the elements within a process, to examine methodically their causal/sequential relationship.

    Such a process may be expressed in the text:

    6.1. in a clear and simple linear fashion

    6.2. not in a clear linear fashion, requiring understanding of a range of cohesive devices to answer the question

7. Ability to identify and synthesise relevant ideas to draw a conclusion

    (The reader draws the conclusion);

8. Transcoding explicitly stated information to diagrammatic display.

9. Understanding grammatical and semantic reference.*

(*This last sub-skill was added to the original list of eight during the rating session)

Appendix A: Reading Subskills in an EAP Test

1. Dealing with relatively uncommon vocabulary:

matching of words/phrases referred to in text with given equivalent meanings.

2. Identification of information in the text, explicitly stated:-

ie, where the question does not paraphrase the text, or where the same key word/ words in both question and text leads the reader to the answer.

The answer may be found in:

2.1. one simple sentence or coordinated sentences

2.2. one complex sentence (judged as complex by factors including subordination, degree of abstractness, negation, verb construction, vocabulary)

2.3. one paragraph or more

3. Identification of information in the text, clearly stated but in paraphrase

(or where no key word occurring in both text and question will lead directly to the answer)

The answer may be found in:

3.1. one simple sentence or coordinated sentences

3.2. one complex sentence (judged as complex by factors including subordination, degree of abstractness, negation, verb construction, vocabulary)

3.3. one paragraph or more

Schooling: A Reader. Cambridge: Cambridge University Press, 196–208.

Weir, C. (1988) 'The specification, realization and validation of an English language proficiency test' in Hughes, A. (ed) *Testing English for University Study: ELT Documents 127*, 45–110.

Yalden, J. (1987) *Principles of Course Design for Language Teaching*. Cambridge: Cambridge University Press.

Lunzer, E., M. Waite & T. Dolan (1979) 'Comprehension and conprehension tests' in Lunzer & Gardner (eds) (1979) *The Effective Use of Reading*. London: Heinemann Educational Books, 37–71.

McNamara, T.F. (1990a) *Assessing the Second Language Proficiency of Health Professionals.* PhD thesis, University of Melbourne.

McNamara, T.F. (1990b) *Item Response Theory and the validation of an ESP test for health professionals.* Language Testing, 7, 1, 52–76.

McDonough, J. (1984) *ESP in Perspective.* London: Collins.

Mead, R. (1982) 'Review of Munby, J. (1978) Communicative Syllabus Design.' *Applied Linguistics*, 3,1, 70–78.

Mossenson, L., P. Hill & G. Masters (1987) *TORCH: Tests of Reading Comprehension. Manual.* Hawthorn, Victoria: Australian Council for Educational Research.

Munby, J. (1978) *Communicative Syllabus Design.* Cambridge: Cambridge University Press.

Oller, J.W. (1975) 'Cloze, discourse and approximations to English' in M.K. Burt and H. Dulay (eds) *New Directions in Second Language Teaching and Bilingual Education: On TESOL '75* . Washington, D.C.: Teachers of English to Speakers of Other Languages (TESOL), 345–355.

Skehan, P. (1984) 'Issues in the testing of English for specific purposes.' *Language Testing*, 1, 2, 202–220.

Spolsky, B. (1988) 'Test review: P.E. Griffin et al. (1986), *Proficiency in English as a second language. (1) The development of an interview test for adult migrants. (2) The administration and creation of a test. (3) An interview test of English as a second language.' Language Testing* 5,1, 120–124.

Tuinman, J. J. (1979) 'Reading is recognition when reading is not reasoning' in Castell, S. de (1986) *Literacy, Society and*