

Testing Reading Comprehension Skills (Part One)

J. Charles Anderson

University of Lancaster

This paper represents an extension of the research reported by Alderson and Luckmani, published in *Reading in a Foreign Language* 5, (2) 1989. On this occasion the research focusses on reading tests from the TEEP and ELTS tests. Judges were presented with (a) taxonomies of reading skills common to construction of these tests and asked to decide whether skills were "High" or "Low"; (b) asked to decide what skills particular items from the TEEP test were in fact testing; (c) asked to decide whether items from the TEEP were testing "high" or "low" skills. In most cases, there was little agreement between judges. Examination of student performance on TEEP items and on items from the ELTS tests showed little relationship between "level" of item and difficulty. It is concluded that we should pay more attention to the processes underlying test performance, and this will be the focus of the second part of the paper, to be published in the next issue of *Reading in a Foreign Language*.

INTRODUCTION

It is commonplace in theories of reading to seek to identify skills which underly or contribute to the reading process. Sometimes the skills identified relate to linguistic features of text in general (see, for example, Munby, 1978), sometimes these skills relate to different sorts of meaning in text, sometimes they relate to supposedly different levels of the understanding that readers can derive from text (see Gray, 1960). Quite frequently, reading skills are explicitly related to Benjamin Bloom's "Taxonomy of Educational Objectives" – see, for example, Adams Smith (1981) which is often taken to represent a hierarchy of skills, such that some skills are thought to be of a "higher order" than other skills, themselves said to be "lower order". For extensive discussions of these skills and related issues, see in particular Clymer (1968), Lunzer, Waite and Dolan (1974), Seddon (1978) and Alderson and Urquhart (1984).

The theoretical nature and status of these skills, and their interrelationships are far from clear, and there are many issues to be resolved, some of which will be referred to in this paper. However, it is the experience of this author that it is common practice among teachers, testers and researchers of reading to assume that reading skills can be identified, taught, tested and researched. It is furthermore common for reading specialists to refer to lower and higher order skills, implying both a hierarchy of such skills, and an implicational scale (or cumulative hierarchy), such that lower skills are held to be necessary before higher ones can be acquired or developed. Note, however, that an implicational scale is not a necessary consequence of a hierarchy – skills may be higher in a hierarchy of values attached to skills. Thus, "the ability to evaluate" may be more highly valued than "the ability to understand explicitly stated information", but may not depend upon the prior

J. Charles ALDERSON is a senior lecturer in the Department of Linguistics, University of Lancaster. His main academic interests are language testing and reading research, and he has published widely in both fields.

possession of such an ability. It remains, however, the impression of the author that most practitioners assume both a hierarchy and an implicational scale. Certainly, Benjamin Bloom considered the six classes of educational objectives (Knowledge, Comprehension, Application, Analysis, Synthesis and Evaluation) to be in a cumulative hierarchy:

Although it is possible to conceive of these major classes in several different arrangements, the present one appears to us to represent something of the hierarchical order of the different classes of objectives. As we have defined them, the objectives in one class are likely to make use of and be built on the behaviors in the preceding class in this list. (Bloom et al 1956, p.18, quoted in Seddon, 1978).

Controversy has surrounded theories of language proficiency for some time, as discussions have centred on whether it can be said to be unidimensional or unifactorial (see, for example, the discussion in Oller, 1983) or whether language proficiency can be said to be multifactorial, and thus consisting of component parts in some set of relationships. The distinction between discrete-point and integrative approaches to test design is long familiar to language testers. Nevertheless, it is common in both discrete point testing and integrative testing to speculate or pronounce on the skills that are being tested by a particular item or test. We will return to this point later in our discussion of the implications of the research to be reported for what it is that tests can be said to be testing.

These two areas of applied linguistic theory – reading and testing – come together when testers, teachers or researchers design a test of reading ability. In such a case, the test designer decides what s/he wants to test – ie what s/he means by reading ability – and finds a means of testing it. After the usual processes of test trialling, analysis, modification and validation (one hopes) one has an instrument in which one can place some confidence, and whose results one can interpret and use to draw conclusions about readers or about reading. Again, whatever the theoretical difficulties or differences of opinion, it is common practice to believe that different aspects of reading – reading skills – are being tested by different tests or items. The research reported below challenges such assumptions and practice.

Seddon (1978) provides a useful overview of attempts to validate Bloom's taxonomy in various educational subject areas. Although his overview challenges many of the research methods, techniques of analysis and interpretations of results of the various studies, his general conclusion is that research has as yet failed to demonstrate convincingly the educational value or psychological reality of the taxonomy. In particular, research shows substantial disagreement among "experts" in assigning test items to particular classes of the taxonomy. Psychometric studies have also failed to demonstrate the posited cumulative hierarchy for objectives. Seddon concludes:

As a final assessment of the validity of the claims concerning the psychological properties of the taxonomy, it is perhaps fairest to say that the picture is uncertain. No one has been able to demonstrate that these properties do not exist. Conversely, no one has been able to demonstrate that they do. (p321)

In the area of reading research specifically, Lunzer, Waite and Dolan (1974) constructed reading tests, for native-speaking children of English, intended to measure different reading skills, but failed to find evidence for an implicational scale among these skills. Nor did they find satisfactory evidence that reading consists of distinct, separately identifiable skills. Instead, they conclude that reading, at least as carried out when taking reading tests, can be considered to be one single aptitude.

Alderson and Lukmani (1989) attempted a partial replication of Lunzer et al. with students for whom English was a second language and had difficulty in getting agreement amongst judges about the skills purportedly being tested by particular items. Furthermore, when item analyses were conducted on those items which judges agreed were testing particular skills, and which they agreed were either lower or higher order skill items, no relationship was found between item difficulty and item level. That is, contrary to expectations, so-called lower order items did not prove to be markedly easier than higher order items – which would be the case if there were an implicational scale. However, when item discrimination was inspected, a relationship between level and discrimination could be discerned, but not the expected one. Contrary to expectations the supposedly lower order items discriminated better among the (ESL) population than did the supposedly higher order items. An inspection of student performances on items showed that the poorer discrimination of “higher order” items came about because students who performed relatively weakly on “lower – order” items performed somewhat better on “higher order” items.

In discussing these results, Alderson and Lukmani speculate that lower order questions might be said to measure language abilities whereas higher order items might measure cognitive skills, reasoning ability, and the like. They conclude that it is inappropriate, therefore, to suppose that students with lower language levels are incapable of answering higher order questions. One should not, therefore, infer from poor performance on lower order questions an inability to perform well on higher order questions. In other words, there is no implicational scale amongst skills, and the pedagogical and testing practices that flow from an assumption of an implicational scale are open to serious questioning.

THE STUDY

This paper reports on subsequent research which has attempted to investigate these issues further, and summarises data presented in Alderson (1987), Alderson and Henning (1987) and Alderson, Henning and Lukmani (1987).

Alderson and Lukmani (1989) produce only very tentative conclusions, since their study was based upon the analysis of the results of one reading test only. Moreover, this test was not standardised, nor is any published evidence available as to its validity. It was therefore decided to partially replicate the research using carefully researched tests for which published data are available.

Two British language proficiency tests – The Test of English for Educational Purposes (TEEP) from the Associated Examining Board, and the ELTS test of the British Council/University of Cambridge Local Examination Syndicate's English Language Testing Service – were the instruments in the study. Both are designed for non-native speakers of English who wish to study at tertiary level in an English-speaking country (usually the UK). Both are explicitly based upon a taxonomic view of language proficiency (Munby, 1978), both contain tests of reading, and both tests can be related directly to explicit statements about what each item tests. For the TEEP, the source of these statements is Weir (1983), for ELTS it is Criper and Davies (1989). Table 1 contains a list of the so-called enabling skills for reading which are tested by the ELTS.

Table 1

Skills being tested on the ELTS test, M1 Study Skills. From Criper and Davies (1988), adapted from Munby (1978) pp176-184.

- | | |
|----|--|
| 15 | Interpreting attitudinal meaning |
| 19 | Deducing the meaning and use of unfamiliar lexical items |
| 20 | Understanding explicitly stated information |
| 22 | Understanding information in the text, not explicitly stated |
| 24 | Understanding conceptual meaning |
| 26 | Understanding the communicative value (function) of sentences |
| 28 | Understanding relations within the sentence |
| 30 | Understanding relations between parts of a text through lexical cohesion devices |
| 32 | Understanding relations between parts of a text through grammatical cohesion devices |
| 34 | Interpreting text by going outside it |
| 35 | Recognising indicators in discourse |
| 39 | Distinguishing the main idea from supporting details |
| 40 | Extracting salient points to summarise |
| 41 | Selective extraction of relevant points from a text |
| 43 | Reducing the text through rejecting redundant or irrelevant information and items |
| 44 | Basic reference skills |
| 46 | Scanning to locate specifically required information |
| 51 | Transcoding information presented in diagrammatic display |
| 52 | Transcoding information in writing to diagrammatic display |

Although Munby (1978) is unclear about the relationships that obtain among these enabling skills, Weir (1983) calls his list of enabling skills an “ordered list” and claims that the enabling skills tested by TEEP are ordered such that lower order skills are required before higher order skills can be deployed (pp 59, 321, 335 and passim)

The methodology of the study was similar to that of Alderson and Lukmani (1989) and consisted of a judgemental and an empirical phase.

THE JUDGEMENTS

In the judgemental phase, 18 experienced teachers of ESL were asked to carry out three tasks.

- (i) After discussion of the posited distinction between higher and lower order skills, and its possible nature, judges were asked individually to rate each enabling skill in Weir’s list for the TEEP as higher or lower (see Table 2)
- (ii) Judges were asked to inspect a reading subtest of the TEEP (TO12 – a 15 item short answer question test on an extended text) and decide which of the 14 enabling skills was being tested by each item. The results are displayed in Table 3.
- (iii) For each test item, judges were further asked to decide whether it was testing a lower or a higher order skill. (see Table 4).

Table 2

Judgements by 18 teachers as to whether the items on the following “Ordered List of Reading comprehension Enabling Skills in an EAP Context” (taken from Weir, 1983) are “higher” (H) or “lower” (L) skills

1 Reference skills, eg using bibliography, index, footnotes	2H	1?	15L
2 Deducing the meaning and use of unfamiliar lexical items through understanding word formation and contextual clues	4L	4?	10H
3 Understanding relations within the sentence	2H	2?	14L
4 Understanding relations between parts of text through grammatical cohesion devices	1H	5?	12L
5 Understanding relations between parts of text by recognising indicators in discourse	6L	2?	10H
6 Understanding the communicative function of sentences, with and without indicators	8L		10H
7 Understanding conceptual meaning, eg, cause, result, purpose	6L	4?	8H
8 Understanding explicitly stated ideas		1?	17L
9 Understanding ideas in a text not explicitly stated		1?	17H
10 Separating essential from non-essential in text: distinguishing the main idea from supporting detail, etc	2L		16H
11 Transfer of information from one medium to another	3L	5?	10H
12 Skimming a) surveying to obtain the gist	4L	3?	11H
b) scanning for specifics	9L	3?	6H
13 Notemaking a) extracting salient points for summary	7L		11H
b) selective extraction of relevant and related points for summary		2?	16H
c) reducing text by rejecting redundant or irrelevant items	3L	1?	14H
14 Critical evaluation		2?	16H

It is clear from Table 2 that there was considerable disagreement among judges on whether particular skills can be said to be of lower or higher order. Unanimity was only achieved on skill 8 – “understanding explicitly stated information” – and in some cases there was an almost even split of opinion (skills 6 and 12 for example).

Table 3

Skills Tested by TEEP TestTO12 as judged by 17 judges (See Table 2 for details of skills)

Item	Skill	Judges																
1.	7/8	3	3	3	3	3/4	5?	4	5	2	2	4	5	3	3	3	2	2
2.	4	4	4	4	4	4	4	4	4	4	4	3	4	5	4	4	3	2
3.	3	3/7	3	3	3	3/4	5?	4	5	2	2	3	5	3	3	3	2	2
4.	6	6/10	8	6	6	6/7	6	13a	8	6	4	6	8	3	6	6	6	8
5.	10	10/13a	8	8	13a	8	10	13b	8	7	4	6	8	10	10	10	8	8
6.	2	2?	8	2	2?	2?	2	2	12b	12	2	2	12b	7	2	2	2?	2
7.	2	2?	8	2	2?	2?	2	2	12b	12	2	2	12b	7	2	2	2	2
8.	8	12b	7	8	8	10	8	8	13a	8	9	8	12a	8	11	8	?	10
9.	8	12b	9	8	8	10	8	9	13a	8	10	8	7	8	11	8	9	1
10.	10	12b	9	8	8	7?	8	10	13a	10	7	8	12	8	10	8	13	1
11.	10	8/12	?10	12b	12a	5	8	?	5	12	13	5	5	10	12a	12a	?	5
12.	1	1	1	1	1	11	11?	1	?	1	9	12	1	1	1	12b	1	1
13.	12b	8/12b	?	13	8	12	8	6	13a	12	7	8	8	12b	8	8	8	?
14.	11	11	9	11	??	11	11	11	11	9	11	8	8	11	11	9	?	11
15.	4/5	5	5	5	4/5	4	8/5	5	4	5	6	5	5	6	5	12a	7/5	?

Table 3 shows the lack of agreement among judges as to which enabling skill was thought to be tested by each test item. The second column indicates which skill the test author believed each item to be testing. The disagreement with the judges is evident, with the possible exceptions of items 2 and 14. In the case of item 1, however, no judge agreed with the test constructor on the skill being tested.

Table 4 shows further lack of agreement among judges as to the level of skill being tested by each item – not surprisingly, given the previous lack of agreement. It should be further noted that there was some evidence of changes of mind as to whether a particular skill is lower or higher order: a person might judge a particular skill to be lower order, but make a different judgement about the level of the skill s/he thought was being tested by a particular item. Since what matters in reading tests is what an item is testing, rather than what its content specifications claim, presumably the data in Tables 3 and 4 is of most interest, and concern.

These results confirm the findings of Alderson and Lukmani (1989) that experienced teachers (different people in the two studies, different tests also) disagree about what most test items are testing. If these judgements are to be believed, either there are serious problems with the tests, or there are problems with the idea that any test item can be said to be testing one particular skill. We will return to this point later, but in either case, there are clearly problems for test validity and validation and conceivably for reading theory.

Table 4

Level of Skills Tested by Items of TEEP TO12
as judged by 16 judges

					Non-native statistics		
					Item No.	Facility Index	Discrimination
	3H	13L			1	0.52	0.32
1L?	1L/H	14L			2	0.58	0.42
		3H	13L		3	0.59	0.45
	9L	3H	3H?	1H/L	4	0.63	0.41
8H	1H?	4L	2L/H	1?	5	0.34	0.45
	1L/H	3H	12L		6	0.53	0.60
	12L	1L/H	3H		7	0.22	0.50
	7H	8L	1L?		8	0.72	0.40
	8H	7L	1L?		9	0.68	0.46
	2H?	6H	8L		10	0.62	0.57
1H?	1H/L	4H	10L		11	0.92	0.35
1?	1H?	2H	12L		12	0.49	0.46
1L?	1L/H	2H	12L		13	0.81	0.55
		2L	14H		14	0.55	0.66
		5L	8H		15	0.38	0.75

TEST DATA

In the empirical phase of the research, again as in Alderson and Lukmani (1989), item analyses were inspected. In the case of the TEEP, item statistics were taken from Weir (1983) where the test validation process is described on a population of native and non-native speakers of English. In the case of the ELTS, the relevant item statistics were taken from Criper and Davies (1988), based upon a population of non-native speakers.

Table 4 (above) shows the item data for the 15 item short answer reading test on TEEP discussed above. The test designer claimed that items 5, 10, 11, 13 and 14 are higher order items. Yet their mean difficulty is .65, notably easier than the lower order items at .53. As in Alderson and Lukmani (1989), the range of difficulties was large, with the most difficult item being a supposedly lower order skill, and the two easiest items being "higher order". With respect to discrimination, however, the findings of Alderson and Lukmani were not replicated, since there appears to be no relation between level of skill and discrimination (for higher order items the range is .35 - .66, mean .52, for lower order items the range is .32 - .75, mean .48).

The possibility exists, of course, that the test designer is "wrong" about the level of skill being tested by each item, and that in those cases where there is relative consensus among judges as to the level of skill an item is testing, that judgement is to be believed. Table 5 presents the item statistics (for both native and non-native populations) for those items where there was relative agreement among judges.

Table 5

TEEP Test Session Test TO12
Items on which judges agree on skills and levels

Items ordered by difficulty

Non-native speakers				Native speakers			
F%	E 1-3	Item No.	Level	F%	E 1-3	Item No.	Level
72	.40	8	L (?H)	96	.34	6	L
68	.46	9	L (?H)	94	.16	2	L
63	.41	4	L	86	.37	9	L?H
58	.42	2	L	83	.64	14	H?
55	.66	14	H?	76	.19	8	L?H
53	.60	6	L	75	.70	15	H
49	.46	12	L	75	.34	4	L
38	.75	15	H	68	.52	12	L

Items ordered by discrimination

Non-native speakers				Native speakers			
F%	E 1-3	Item No.	Level	F%	E 1-3	Item No.	Level
38	.75	15	H	75	.70	15	H
55	.66	14	H?	83	.64	14	H?
53	.60	6	L	68	.52	12	L
68	.46	9	L?H	86	.37	9	L (?H)
49	.46	12	L	96	.34	6	L
58	.42	2	L	75	.34	4	L
63	.41	4	L	76	.19	8	L (?H)
72	.40	8	L?H	94	.16	2	L

Unfortunately, there are very few "higher order" items on which judges agree. However, the previous result does seem to be confirmed, that there is no relationship between difficulty and level of skill, both for native and for non-native speakers. There is, however, a suggestion that higher order items might discriminate better than lower order items, for both populations. As these items – numbers 14 and 15 – also happened to be the last two items on this test, this could be a speed effect, rather than the effect of item level.

In order to investigate this issue more carefully, it was necessary to look at a larger sample of items than those available on test TO12. Table 6 shows the item statistics for seven tests of reading from the two parts of the TEEP, and data from a multiple choice grammar test for comparison. Items have been categorised as higher or lower, as indicated by Weir (1983), and the results summed.

Table 6

Average Item Statistics for TEEP Test Sessions 1, 2A and 2B
Native and non-native speakers

Data taken from Weir (1983) and categorised as to Higher or Lower Order Skills
(Skills 1 - 8 = Lower Skills; 9 - 14 = Higher)

	Mean Difficulty	Mean Discrimination	n items
Test TO12 (Short Answer questions)			
Non-natives (n=330)			
High	65%	.52	5
Low	53%	.48	10
Natives (n=125)			
High	81%	.38	5
Low	83%	.38	10
Test TA12 (Gap-filling)			
Non-natives (N=439)			
Low	40%	.54	21
Natives (N=72)			
Low	78%	.41	21
Test TB12 (Gap-filling)			
Non-natives (n=326)			
Low	48%	.52	21
Natives (n=125)			
Low	83%	.38	21
Test TA11 (Multiple-choice questions)			
Non-natives (n=438)			
High	41%	.48	5
Low	68%	.33	8
Natives (n=71)			
High	68%	.48	5
Low	86%	.29	8
Test TB11 (Multiple-choice questions)			
Non-natives (n=321)			
High	68%	.52	4
Low	72%	.37	10
Natives (N=124)			
High	80%	.36	4
Low	84%	.24	10
Test TA13 (Short answer questions)			
Non-natives (n=435)			
High	73%	.49	6
Low	62%	.46	6
Natives (n=76)			
High	90%	.31	6
Low	80%	.32	6

Table 6 (cont.)

Test B13 (Short answer questions)			
Non-natives (n=325)			
High	30%	.54	5
Low	60%	.41	5
Natives (n=125)			
High	67%	.48	5
Low	84%	.36	5
Test TO41 (Multiple choice grammar)			
Non-natives (n=333)			
Low	69%	.41	60
Natives (n=133)			
Low	95%	.25	60

Once again, the finding is confirmed that there is no relationship between level of skill and item difficulty. Sometimes (test TO12, TA13) "higher order" items are easier, sometimes (test TA11, TB13) they are more difficult. Whatever causes item difficulty is not related to the level or difficulty or complexity of the skills, if items are measuring identifiably separate skills.

With respect to item discrimination, however, the Alderson and Lukmani results are clearly not replicated. In every case where comparative data is available, "higher order" items discriminate better than lower order items.

Before reaching any conclusions about this disparity, however, it seemed sensible to investigate whether different test constructors and test analysts might produce clearer results. Therefore, item data was taken for the ELTS test from Criper and Davies (1988). In that study, the authors claim that particular skills can be said to be tested by particular items. Although they provide no evidence for their assertion, and no data is available on pooled judgements of informants, it was decided provisionally to accept their claim, in order to test the hypothesis that item difficulty and item discrimination are not related to the level of the skills being tested. Again, the assumption was made that it was possible to categorise the items, based upon the description of their enabling skills, as being either "higher" or "lower" in level. Because of some doubts about the relevance to reading of the Munby skill 44 – "basic reference skills" – items supposedly testing this were separately categorised. Items were grouped and the statistics are presented below for both a general Reading test – G1 – and a subject-specific reading test in particular discipline areas (M1 in six disciplines).

Table 7 reveals a clear lack of relationship in both difficulty and discrimination between "higher" and "lower" order items. In three tests, "higher" order items are more difficult, in three tests, there is no difference, and in one test higher order items are easier. In three tests, higher order items discriminate less well than lower order items; in four tests, they discriminate better.

Table 7

Average Item Statistics for ELTS Test
M1 – Study Skills – and G1 – Reading

Data taken from Chriper and Davies (1988) and categorised as Higher and Lower Order Skills

	Mean Difficulty	Mean Discrimination	n items
Test M1 Study Skills			
General Academic (n=359)			
High	49%	.44	18
Low	54%	.50	12
Reference	42%	.62	5
Medical Science (n=122)			
High	61%	.46	7
Low	66%	.41	17
Reference	60%	.43	4
Physical Science (n=137)			
High	63%	.55	8
Low	62%	.45	6
Reference	79%	.45	4
Technology (n=168)			
High	75%	.28	5
Low	74%	.33	17
Reference	72%	.44	4
Social Science (n=238)			
High	50%	.40	13
Low	56%	.36	21
Reference	48%	.36	3
Life Science (n=301)			
High	63%	.38	4
Low	54%	.43	10
Reference	44%	.39	6
G1 – Reading (n=1325)			
High	72%	.45	4
Low	71%	.37	34

If we compare these results with the apparent contradiction between the Alderson and Lukmani (1989) findings, and the above findings, and the above findings from the TEEP test, it seems not unreasonable to conclude either that there is no stable relationship between “level” of skill being tested by items, and the items’ difficulty and discrimination, or that the items are not clearly distinguishable by “level”.

DISCUSSION

The above findings suggest the following preliminary conclusions:

- i) Judges are unable to agree as to what an item is testing.
- ii) Judges are unable to agree upon the assigning of a particular skill to a particular test item.
- iii) Judges are unable to agree upon the level of a particular skill or a particular item
- iv) There appears to be a lack of relationship between item statistics and what an item is claimed to be testing

There may be several explanations for these findings.

- a) There are serious reasons for doubting whether a skill can be said to be "higher" or "lower" than another skill in any hierarchy that implies relative difficulty or some differential stage of acquisition (at least for ESL readers)
- b) The skills identified by Weir (1983) and Davies and Cripser (1988), both based on Munby (1978), are in fact overlapping rather than discrete.
- c) It is inappropriate to categorise the skills in the Munby taxonomy into "higher" and "lower" order skills.
- d) It is unlikely that any test item can be unambiguously said to be testing any one skill.

The final point above surely accords with the common-sense view that answering a test question is likely to involve a variety of interrelated skills, rather than one skill only, or even mainly. Even if there are separable skills in the reading process which one could identify by a rational process of analysis of one's own reading behaviour, it appears to be extremely difficult if not impossible to isolate them for the sake of testing or research.

However, this conclusion has to be seen in the light of the following consideration from Alderson and Lukmani (1989):

It is likely that the manner in which individuals arrive at any answer to a question will vary. One person may have difficulty with a particular word and need to infer connections across sentences. Another may understand the word and, therefore, not need to infer. Thus a product, that is, a right answer, may be arrived at in a variety of different ways using different processes, strategies or skills. (p.264)

It is interesting to note that Bloom et al (1956) suggest that, depending on the nature of prior learning experiences, different students can solve the same question in different ways, and that one objective or test item can be placed in different categories (p.16). However, as it is unlikely that any two students can be said to

have the same learning experiences, it is curious that Bloom and subsequent researchers have failed to address this central issue.

There are significant implications of this insight, if true, for reading research and testing. Much reading research uses tests of comprehension in order to make inferences about reading ability, reading skills and reading processes. Inferences are made on the basis of the answer a test taker has chosen. If the relationship between process and product is not only indirect, but also variable across individuals, such inferences are unjustified, and therefore so are many of the conclusions and theories that have been built upon test results.

Testing builds upon the assumption that it is possible to discover what ability a test or an item measures. Such discoveries depend upon making inferences about what a test-taker is doing when answering a question or responding to an item. If this varies across individuals then the validity of a test must also be in question. Furthermore, if judges cannot agree on what an item or test is testing – perhaps because they themselves arrive at correct answers in a variety of different ways – then we are apparently forced to abandon a judgemental approach to content validation and do one of two things: either to make inferences about test validity from test results, and their statistical relationships internally and externally (which implies inferring process from product) or to examine much more carefully the processes that test takers go through when responding to test items.

Recent years have seen the beginnings of research into the processes that test takers go through during a test. It should prove possible to illuminate, through introspective and retrospective accounts from test takers, the issue of whether reading test items designed to test particular skills do indeed measure such skills (whether they are ordered in a cumulative hierarchy or in some other arrangement with a taxonomy). It should also prove possible to examine the relationships among such skills through the same means. The second part of this paper will report on a pilot attempt to do precisely this.

BIBLIOGRAPHY

- Adams-Smith, D. (1981) "Levels of Questioning: Teaching Creative Thinking through ESP" *Forum*, Vol 19, no 1 pp 15-21
- Alderson, J Charles (1987) "Testing Reading Comprehension; The Notion of Hierarchically Ordered Skills" Plenary address, First International Conference on Language Testing, Tsukuba University, Tokyo, Japan, March 1987.
- Alderson, J. Charles and Grant Henning (1987) "Testing Reading" Paper given at the Conference on Developments in Language Testing Research, University of Reading, UK, April, 1987.

Alderson, J. Charles, Grant Henning and Yasmeeen Lukmani (1987) "Levels of Understanding in Reading Comprehension Tests" Paper presented at the Ninth Annual Language Testing Research Colloquium, April 1987, Miami, USA.

Alderson, J. Charles and Yasmeeen Lukmani (1989) "Cognition and Reading: Cognitive Levels as Embodied in Test Questions" *Reading in a Foreign Language*, Vol 5, no 2 pp 253-270

Alderson, J. Charles and A.H. Urquhart (1984) *Reading in a Foreign Language*. Longman.

Bloom, B.S., Englehart, M.D., Furst, E.J., Hill, W.H. and Krathwohl, D.R. (1956) *Taxonomy of Educational Objectives. Handbook I: Cognitive domain*. New York: Longmans, Green

Clymer, T. (1968) "What is reading?: some current concepts" Reprinted in Melnik, A and J Merritt (1972) *Reading Today and Tomorrow* University of London Press, pp 48-66.

Criper, C. and A. Davies (eds) (1988) *ELTS Validation Project Report*. English Language Testing Service Research Report Vol 1(i). The British Council and The University of Cambridge Local Examinations Syndicate.

Gray, (1960) "The Major Aspects of Reading" in Robinson, H (ed) *Sequential Development of Reading Abilities* Supplementary Educational Monographs, no 90. Chicago University Press, pp 8-24.

Lunzer, E. Waite, M. and Dolan, T. (1979) "Comprehension and Comprehension Tests" in Lunzer, E. and Gardner, K. (eds) *The Effective Use of Reading* Heinemann Educational Books, pp 37-71.

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

Oller, J. W. (ed) (1983) *Issues in Language Testing Research* Rowley, Mass: Newbury House.

Seddon, G. M. (1978) The Properties of Bloom's Taxonomy of Educational Objectives for the Cognitive Domain. *Review of Educational Research*, Vol 48, no 2, pp 303-323.

Weir, C. J. (1983) *Identifying the Language Problems of Overseas Students in Tertiary Education in the UK*. Unpublished PhD thesis, Institute of Education, University of London.