DOCUMENT RESUME

ED 322 753 FL 018 764

AUTHOR Alderson, J. Charles
TITLE Judgements in Language Testing, Version Three.
PUB DATE Apr 90
NOTE 13p.; Paper presented at the Meeting of the World Congress of Applied Linguistics (9th, Thessaloniki, Greece, April 15-21, 1990).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; Difficulty Level; *Evaluative Thinking; *Item Analysis; Language Research; *Language Tests; *Testing; *Test Items; *Test Results

ABSTRACT

Language testing is an area of applied linguistics that combines the exercise of professional judgment about language, learning, and the nature of the achievement of language learning with empirical data about student performance and, by inference, their abilities. The relationship between judgments and empirical data in language testing is examined through three studies. The first investigates language professionals' judgments about test content and the skills and abilities supposedly being tested by certain test items and compares them with test results and the attitudes of test-takers. The second study compares the judgments that experienced test writers and scorers make about item and test difficulty and compares them with item results. The third study gathers judgments from language testers and teachers about standards of performance of a given population in a standard-setting exercise aimed at determining grade boundaries for a public examination. The implications of the findings are discussed. (Author/MSE)

Judgements in Language Testing

Version Three

J Charles Alderson,
University of Lancaster

Abstract

Language testing is an area of applied linguistics that combines the exercise of professional judgement about language, learning and the nature of the achievement of language learning, with empirical data about students' performances and, by inference, their abilities. This paper addresses the issue of the relationship between judgements and empirical data in language testing by reporting on three studies. The first study investigates judgements by language professionals of test content and the skills and abilities supposedly being tested by certain test items, and compares these judgements with the results of test administrations and the introspections of test-takers. The second study compares the judgements that experienced test writers and examination markers make about the difficulty of items and tests, and compares those judgements with the results of an administration of the items. The third study gathers judgements from language testers and teachers about standards of performance of a given population, in a so-called standard setting exercise aimed at determining grade boundaries for a public examination. The paper ends by discussing the value of professional judgements in determining test content, test difficulty and criterial cut-offs.

Introduction

Applied linguists are frequently called upon to make professional judgements. They may be asked to comment upon the appropriacy of a set of pedagogic materials for their avowed purpose; they may be invited to judge the merits of competing syllabuses, or approaches to language teaching. They may be asked to decide the extent to which a language teaching theory is adequately reflected or operationalised in classroom practice, or to reflect upon the relationship between a particular view of language competence and the content of a textbook.

Language testers, as applied linguists, are frequently required to make similar sorts of judgements about test content, test method, and their appropriacy for a given purpose, or the extent to which they reflect particular theoretical approaches. Indeed, language testing is an area of applied linguistics where judgements are needed at every level of activity and every stage in test development and validation. Testers have to judge whether test specifications are fit for their purpose, whether test content reflects the test's specifications, whether the test method is appropriate for the test's purpose, whether scoring criteria are appropriate and whether candidates' performances meet those criteria. Judgements abound in language testing, and are inevitable, even, as Albert Pilliner pointed out in 1968, in so-called "objective" tests.

Language testing is characterised by its attempts, indeed its determination, to corroborate judgements in some areas - as in the case of making subjective judgements on candidates' performances, where it is usual practice to estimate the intra- and inter-rater reliability of judgements - or, in other cases, to substitute

empirical data for judgement - as, for example, in the common practice of pilotting test items in order to determine their difficulty levels, rather than relying upon the test constructor's judgements of that difficulty level. However, whereas in language testing research it is common to meet studies of the reliability of scorer judgements - the inter- and intra-rater reliability studies referred to above - it is unusual to find studies of other aspects of the judgements that language test constructors have to make.

This paper attempts to add to our understanding of the nature of judgements in language testing by examining under-researched areas where judgements are important. It will be argued, and hopefully demonstrated, that so-called professional judgements are frequently flawed, or in serious conflict with other professionals' judgements, and that language testers must be alert to the need to corroborate and validate the professional judgements we so frequently make. The exercise of judgement, and the variability of those judgements, will be illustrated in three areas: test content, test difficulty, and determining test cut-offs.

Study One: Test Content

Part of the process of test construction involves turning intentions into actions: in other words, having decided what they want to test, test constructors have to produce items that embody those intentions. Test specifications have to be translated into test items, and an important part of the judgemental process in testing is deciding whether specifications have been adequately realised: whether a test item does indeed test what the specifications say it tests.

Alderson and Lukmani (1989) report on a study in which 10 native speaker judges were asked to decide what skills were being tested by items in a reading test, and whether the skills being tested were higher or lower order reading skills. In over half the items being judged, there was little or no agreement on which skills were being tested, or what "level" of skill was involved. For example, Alderson and Lukmani report some judges claiming that a given item tested the skill of vocabulary recognition, others the ability to distinguish and discriminate elements in a context, others the ability to break entities down into their constituent parts and to perceive relationships, yet others the ability to clarify the meaning of complex ideas, and still others the ability to synthesise and recognise patterns.

For one item the descriptions given by the ten judges were:

Judge 1: Recognition of vocabulary and grammar
Judge 2: The words "where", "bottom", "Atlantic"; knowledge that settings are early in a narrative
Judge 3: Scan for specifics, relations in a sentence
Judge 4: Detailed information, recognise prepositions, links between parts of complex
Judge 5: Word meaning, text organisation, accessibility of "sea-bed" schema
Judge 6: Inferred meaning; discourse rules re mentions
Judge 7: Direct verbatim extraction of information
Judge 8: Establish context, lift a phrase direct from text
Judge 9: Direct reference, intersentence connections
Judge 10: Understand text links (from Alderson and Lukmani (1989 p263)

The authors concluded that judges had difficulty in agreeing on the skills being tested by particular test items. They recognised, moreover, that there was likely to be a degree of intra-rater unreliability in the judgements, such that if the judges had been asked to repeat their judgements after some time lapse, there might well have been differences in the descriptions assigned to particular items.

In a subsequent study (Alderson, 1988) 17 judges were asked to identify which skill from a list of 14 skills (adapted by Weir, 1983 from Munby, 1978) was being tested by each item in a 15-item reading test designed by Weir (1983). For each item, the judges were further asked to decide whether the skill being tested was a "higher" or a "lower" order skill. It was clear from the results that there was again considerable disagreement among judges as to which so-called "enabling skill" was being tested by each test item. Indeed, in one case, no judge agreed with the test constructor on the skill being tested by that item; and in no case was there complete agreement, either among judges, or between the judges and the test constructor, as to the skill being tested.

There was further lack of agreement among judges as to the level of skill being tested by each item. It was also noted that there was some evidence of changes of mind as to whether a particular skill was "lower" or "higher" order: a judge might say that a particular skill as described in the test specifications was "lower" order, yet make a different judgement about the level of the actual item that he thought tested that particular skill.

Alderson took these results to confirm the findings of Alderson and Lukmani (1989) and concluded that:

"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"

Furthermore, after examining the relationship between the judgements of the judges, and the empirical item difficulties and discriminations, he further concluded:

"iv) There appears to be a lack of relationship between item statistics and what an item is claimed to be testing."

Alderson then makes the following comment:

"Testing builds upon the assumption that it is possible to discover what ability a test or an item measures. ... 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.... or to examine more carefully the processes that test takers go through when responding to test items. However, if, as seems prima facie likely, different readers do different things, what can a test item be said to be testing?"

In the second part of his study, Alderson (1988) provided tentative evidence from studies of test-takers' introspections that individuals did indeed approach test items in different ways, revealing that they were engaging in a variety of different skills simultaneously and in an overlapping manner. He suggested that the evidence for separable skills was not overwhelming, and that therefore test specifications might need to be rewritten. The evidence suggested that the test constructor was wrong in his characterisation of what an item was testing, and also that judges, insofar as they disagreed about what an item was testing, were probably also wrong, or at best only partially right, in their judgements.

"What a test of reading tests is not simply what its constructors say it tests, nor what a set of judges consider it to test. It must surely and crucially relate to what happens inside a test-taker's head when he or she responds to an item. Finding out that information, and discovering how generalisable the results are, is a neglected but important research endeavour."

Study Two: Test and Item Difficulty

In order to examine whether experienced teachers, examiners and test constructors could accurately predict item and test difficulty from an inspection of items, Alderson and Rajapakse (forthcoming) enlisted the cooperation of 21 Sri Lankan teachers of English with varying degrees of familiarity with and responsibility for a new National Certificate in English test in Sri Lanka. This new test had been introduced into Sri Lanka by the Ministry of Education in 1986: it was intended to provide a prestigious qualification of proficiency in English, for adults outside the formal education system who wished to enhance their employment prospects through producing evidence of their level of English. (The new test was also intended to provide the opportunity to experiment with innovative testing methods, prior to the introduction of examination reform within the secondary school system.) The new test was designed by a restricted number of teachers on the staff of the Curriculum Development Division of the Ministry, and the examinees' performances were scored by raters trained and supervised by Team Leaders. The 21 teachers who took part in the judgemental task included: three highly experienced testers who had been responsible for designing the test paper, and also for training markers, and for supervising marking; 5 testers who had written some test items, and taken part in the supervision of markers; 5 testers who had been Team Leaders during the marking; 5 minimally trained testers who had been involved in marking the test as part of a team, and 3 testers with no experience of writing items for the NCE, nor of marking the 1986 paper.

Several weeks after the examination paper had been marked, and the testers had returned to their normal duties, they were given a copy of the 1986 test paper and asked to indicate, against each item, what approximate percentage of the candidates taking the examination they thought would have answered that item correctly. (The test paper consisted of 30 Reading items across five subtests, 75 C-test items on 3 passages, and 60 cloze test items on 3 passages.) The estimated p values for each item could then be totalled to arrive at estimated sub-test difficulties.

It was thus possible to compare estimated difficulties with the actual difficulties of the items and sub-tests. The hypothesis was that experienced test designers and examiners would be able to predict item and test difficulty in advance of a test administration. In that case, there would be no need for pre-testing of test items - a finding which would have greatly convenienced the system, since item pre-testing was virtually unknown in Sri Lanka.

Table One presents the correlation matrix between the empirical item difficulties, and the judgements of the 21 judges.

Table One

Correlations between judges' estimates and actual item difficulties

Judge/	Reading	WC1	WC2	WC3	CL1	CL2	CL3	OVERALL
1	.38	.67	.60	.41	.58	NS	NS	.49
2	NS	.51	.52	.52	.51	NS	NS	.49
3	.71	.35	.56	.57	.52	NS	.51	.64
4	.61	NS	NS	NS	NS	.46	NS	.52
5	.70	.47	NS	.65	.52	NS	.75	.58
6	.49	.36	.56	.37	NS	.39	.47	.54
7	.35	.44	NS	NS	.50	NS	NS	.51
8	.68	NS	.58	.48	NS	NS	NS	.37
9	.68	NS	.34	.44	NS	NS	NS	.28
10	.32	.39	NS	.42	.52	NS	NS	.33
11	.63	NS	-.48	NS	NS	NS	NS	.36
12	.74	.57	.61	.62	NS	.45	NS	.64
13	.63	.54	NS	NS	NS	NS	NS	.52
14	.69	NS	NS	.57	.43	NS	NS	.45
15	.75	.48	.41	.65	NS	.48	NS	.59
16	.67	.63	.51	.44	NS	NS	NS	.56
17	.68	.51	.34	.54	.81	NS	NS	.45
18	.52	.53	.74	.49	.60	NS	NS	.41
19	.77	NS	.47	NS	.43	NS	.43	.61
20	NS	.55	.50	.48	.52	.53	NS	.52
21	.49	-.40	.53	NS	-.30	NS	NS	.27

Table One shows clearly that there was no marked agreement between the judges and the empirical difficulties. Nor is there any discernible tendency for highly experienced testers to predict difficulty better than inexperienced testers (judges 1- 3 compared with judges 19 - 21). Although the experienced judges fare slightly better, and achieve fewer very low coefficients, they are still far from accurate in their predictions. Table Two below summarises this:

Table Two
Highly experienced vs inexperienced judges

	Reading	WC1	WC2	WC3	CL1	CL2	CL3
Exper.	.79	.63	.71	.61	.68	NS	.45
Inexp.	.67	NS	.71	.47	.51	NS	NS

In general, it appears to be more difficult to predict the difficulties of the C-tests and the cloze tests than those of the Reading tests. Yet although most judges (regardless of experience) seem better at predicting Reading test difficulty, there is no obvious reason why this should be the case, and they are still far from perfect.

Tables Three and Four below present the actual subtest difficulties, compared with the mean item difficulties for all judges, and for experienced versus inexperienced judges.

Table Three
Actual and judged test difficulties (in %)

	Reading	WC1	WC2	WC3	CL1	CL2	CL3	OVERALL
Test mean	**50**	**48**	**27**	**30**	**28**	**12**	**27**	**33**
sd	**23.9**	**18.9**	**18.6**	**22.7**	**22.1**	**11.6**	**18.8**	**23.5**
Judge 1	67	67	57	60	64	52	59	61
Judge 2	47	35	29	24	16	10	36	29
Judge 3	57	38	28	32	15	22	21	33
Judge 4	56	53	39	35	20	24	23	38
Judge 5	62	46	23	48	37	32	33	41
Judge 6	71	50	44	54	36	29	54	50
Judge 7	69	50	38	34	11	11	10	34
Judge 8	64	86	65	81	82	67	64	73
Judge 9	61	49	40	43	50	61	45	50
Judge 10	62	25	12	18	20	35	36	31
Judge 11	46	8	9	10	21	14	8	18
Judge 12	64	68	44	42	52	23	23	47
Judge 13	67	66	51	43	58	22	20	48
Judge 14	66	56	40	62	53	43	61	55
Judge 15	61	46	37	53	46	34	38	46
Judge 16	54	60	39	20	36	22	31	38
Judge 17	71	91	80	88	62	52	56	72
Judge 18	62	44	41	40	49	50	56	49
Judge 19	84	77	26	31	47	22	25	47
Judge 20	73	58	40	48	43	41	43	51
Judge 21	64	62	62	49	55	53	55	57
Overall	63	54	40	43	42	34	38	46
sd	8.7	19.2	16.7	19.3	18.6	16.7	17.1	13.6

Table Four
Experienced versus inexperienced judges mean judgements

	Reading	WC1	WC2	WC3	CL1	CL2	CL3
Exper.	57	46	38	39	33	28	39
Inexper.	74	66	43	42	48	39	41

Whatever the patterns that may be (faintly) discernible in the data, and whatever tentative reasons might be advanced to account for such patterns, it is clearly the case from this study that testers, both very experienced and inexperienced, have problems in predicting item and test difficulties. These results justify the recommendation that all test items should be pre-tested before incorporation into new versions of the NCE in Sri Lanka, on the grounds that the judgements of experienced testers about likely item difficulty are simply too variable and inaccurate to be trustworthy.

Study Three: Deciding on Grade Boundaries

As part of a research project intended to evaluate a new school-leaving examination in English in a country whose identity must remain confidential, the principal researchers, of whom the current author was one, investigated ways in which cut-off points, or grade boundaries, could be determined for the new examination (results had to be reported as Pass, Credit, Distinction and Fail). This examination represented a radical departure in both content and method from the previous examination, and it was therefore necessary to advise the relevant authorities on how examination grades might be established. (The operational circumstances were such that it was impossible to administer both new and old examinations to a sample of students in order to determine the comparability of examinations.)

In an attempt to get some notion of criterion-referencing into the decision making process with respect to grade boundaries for the examination, a small study was undertaken. 19 individuals were given a copy of the new examination papers (Papers One and Two), and were asked to make three sets of judgements. The first set - referred to as Global judgements - related to what marks specified stereotypical candidates would gain on each paper. The instructions were as follows:

> "Think of pupils you have taught. Think of people who you consider to be just barely a Pass at O-Level English Language. Call them "Bare Pass".
>
> Think of other pupils who you consider to be only just barely a Credit at O-Level English Language. Call them "Bare Credit".
>
> Think of a third group of pupils who you consider to be just barely a distinction in O Level English Language. Call them "Bare Distinction"."

For each category of candidates, judges were asked to consider what mark would be achieved on Paper One (out of 60) and Paper Two (out of 140). These judgements were called "Global" judgements.

In the second set of judgements, for the same categories of candidates, judges were asked to decide the percentage of candidates

who would get each question correct, or, in the case of the Writing questions, what score (out of 5 or 10, 15 or 20, depending on the question) each category of candidates would achieve. These item level judgements were then summed to give an "Item-level Total" judgement.

Finally, judges were also asked to say what overall total score (out of 100%) each of the three categories of pupils would get. These judgements were labelled "Overall Total".

Judges included experienced and trained language testers, who were closely involved in the development of the new examination, and experienced teachers of English at prestigious schools in the country. The details of the experience and expertise of individuals are unknown but this must have varied.

The results showed considerable variation in the judgements made, both at Global and at Item levels. It was, however, possible to pool judgements across 19 judges in the case of the Global judgements and 14 judges in the case of Item-level judgements. (In the latter case, obvious inconsistencies were omitted, for example where a judge indicated that a Bare Pass pupil would get a higher score for an item than a Bare Credit pupil would. In such cases, it was assumed that some misunderstanding had occurred with respect to the instructions and the judge was omitted from the analyses.)

The results were as follows:

Table Five

Percentage marks judged appropriate for particular grade boundaries

	PASS	CREDIT	DISTINCTION
GLOBAL	34%	51%	71%
ITEM-LEVEL	26%	51%	75%
OVERALL TOTAL	30%	48%	68%
TRADITIONAL	30%	50%	75%

The "Traditional" percentage marks are the cut-offs believed to have been used by the relevant authorities in determining grade boundaries. It should be pointed out that these percentages do not appear to have varied with the difficulty of the examination.

Whatever variation there might have been in individual judgements, there is a considerable degree of agreement among the pooled judgements arrived at in the three different ways, when compared to the "Traditional" cut-offs. It is obvious from Table Five that the percentage cut-off judged to be appropriate, even when judgements are pooled, varies according to the method used to collect the judgements. The pooled Item-level judgements result in an "underestimate" of the cut-off for Bare-Pass, and pooled Global judgements result in an "overestimate" of the cut-off for the same candidates. (However, it should be remembered that no independent data is available on what the

cut-offs "should" be.)

Having identified the possible cut-offs, it was then possible to calculate the proportion of candidates gaining each grade for each of these putative decisions, to compare effects. Calculations were based upon the distributions of the total population (c. 200,000).

The consequences of using each set of cut-offs are given in Table Six below. For comparative purposes, the actual proportions of candidates gaining each grade in three previous years is also given.

Table Six

Proportion of the population gaining each grade

	PASS	CREDIT	DISTINCTION	FAIL
GLOBAL	18.1%	7.1%	2.6%	72.2%
ITEM-LEVEL	35.8%	7.5%	2.2%	54.5%
OVERALL TOTAL	22.1%	8.1%	3.4%	66.4%
1987	28.6%	5.9%	2.1%	63.4%
1986	28.7%	8.8%	2.1%	60.4%
1985	15.2%	5.0%	1.4%	78.4%
HISTORICAL AVERAGE	23.1%	8.3%	2.2%	66.4%

From the above, it is obvious that considerable variation in proportions passing, failing, etc has taken place over the years. This is obviously because of the rigid application of fixed grade boundaries - assumed to be 30%, 50% and 75% respectively - without regard for the variation in difficulty or ease of the actual examinations set. (In such a large population, it is highly unlikely that levels of achievement vary much from year to year, at least until the recent provision of new textbooks and teacher training can be supposed or shown to have had some impact.)

It is also obvious from Table Six that the percentage of students failing or gaining a bare pass in the new examination will vary considerably depending upon the method used to determine the appropriate cut-off mark.

Clearly, a decision on grade boundaries is a complex matter, and the possibilities are numerous. One can decide to award grades on a proportional basis, so that roughly the same percentage of candidates gets a given grade each year, unless some change in circumstances warrants otherwise. This has obviously not been the practice to date.

Alternatively, one can decide to have the same percentage marks as boundaries as in previous years (ie 30, 50, 75). That would give, as

indicated above, a failure rate for the new examination of some 66%. However, such a practice is theoretically only justified if one has reason to believe that the current examination is equivalent in difficulty and reliability to previous exams, something which has not been established, and which is fairly unlikely - as, indeed, is shown by the fluctuation over time in proportions of candidates being awarded any given grade.

Another alternative is to take the judgements of experienced teachers and testers, who might be expected in the light of their experience of teaching and of marking exams over a period of time to have arrived at an internalised notion of difficulty in relation to candidate ability. Interestingly, the OVERALL TOTAL judgements would give broadly similar results to the average Historical grade boundaries (66% failure, 22% passes, 8% credit and 3.4% distinctions, although OVERALL TOTAL judgements result in a Distinction cut-off mark of 68, rather than the historical 75).

In the absence of data comparing the difficulty of the 1988 exam to that of previous exams, it seems reasonable to assume that, on the whole, candidate abilities have not changed greatly from previous years. Three possible recommendations follow from these discussions:

1. That all future trialling of draft examination items be done alongside equivalent or comparable items from the previous year's exam paper, to enable direct comparisons of the difficulty of papers.

2. That, for the new 1988 examination, the grade boundaries of 30, 50 and 75 be retained. This would give a failure rate of 66% - higher than 1986 or 1987, but not by much, and considerably lower than 1981 - 1985. This might be felt to be justified in the light of the 1988 exam's improved quality and relevance to the use of English.

3. That grade boundaries not be seen as fixed for all time, but as subject to change depending upon the performance of candidates and the difficulty of future items and exams as a whole.

It should, however, be stressed that this research project was only intended to throw some light upon ways in which judgements could be gathered that might inform decisions on grade boundaries, and on the consequences that might flow from using one method rather than another to arrive at possible cut-offs. Whatever decisions were eventually reached by the appropriate authorities are unknown to the present author, and such decisions were, of course, taken completely independently of any "outside" advice.

Despite the differences among judges in this standard-setting exercise it was felt that gathering judgements had had some value, in that the results of the pooling of the judgements had not resulted in wildly different results from decisions that would have been made using traditional criteria. In this particular circumstance, corroborating data were not available: there were no candidates who had taken the new exam and the previous exam, and although there were item and test difficulty statistics, there was no independent evidence of what constituted an adequate performance at each boundary. It is thus arguable that one has to place considerable faith in the quality of the pooled judgements, whatever the variation in individual judgements.

Summary and Conclusion

This paper has examined the results of three studies of judgements. In the first study, considerable disagreement was found among judges as to what items in reading tests were testing. Even when agreement was discernible, what the judges claimed an item to be testing did not necessarily agree with the test constructor. Nor did their claims relate in any predictable way to item difficulty and discriminations. It was seen to be necessary at least to corroborate judgements about test content by investigating what test-takers reported about their performances and their reasons for the decisions they made in responding to test items.

In the second study, judges were unable to predict with any degree of accuracy the difficulty of test items and sub-tests. As a result, the need was established for pre-testing of items and sub-tests, independently of the opinions of test constructors as to item difficulty.

In the third study, there was considerable variation in the judgements gathered, and even the pooling of judgements resulted in different recommended cut-offs, depending upon the method used to gather the judgements.

In the latter study, the judgements gathered could be accepted, with some caution, in the absence of other data on the appropriacy of cut-offs. In the second study, the judgements had to be rejected as a basis for decisions about items. In the first case, it was necessary to have recourse to different data sources in order to get a clearer picture of what was being tested.

In the case of the first study, however, it was admitted that some of the variation in judgements could have been caused by inadequate test specifications, or models of the reading process. That is, the assumption is likely to have been wrong that one can isolate and test "enabling skills" separately. Rather, it is likely to be the case that a variety of skills are used in some as-yet-not-understood fashion (but probably in an integrated way) when answering reading test items. Thus, gathering judgements may have led to further insights into test design, and to the need for a revision of test specifications.

Similarly in the second study, the comparison of judgements with empirical difficulties highlights the need, not only for pre-testing of items to establish their difficulty levels, but also for the provision of feedback to item writers, in an attempt to train them to enable them to make more accurate predictions of item difficulty.

The third study reveals the difficulty of making criterion referenced cut-offs for grade boundaries, and highlights the need for further research in the area.

All three studies, however, present evidence that shows both the usefulness and the limitations of gathering judgements from professional testers. There is a clear need, in all cases where judgements are involved, for corroboration of the accuracy of the judgements before accepting that they are necessarily valid. We need to ensure that all our judgements, not just the judgements of markers, are open to scrutiny and challenge.

Bibliography

Alderson, J Charles (1988) "Testing Reading Comprehension Skills" Paper presented at the Sixth Colloquium on Research in Reading in a Second Language, TESOL, Chicago

Alderson, J Charles and Y Lukmani (1989) "Cognition and Reading: Cognitive Levels as Embodied in Test Questions" Journal of Reading in a Foreign Language, Volume 5, No 2, p253-270

Alderson, J Charles and L Rajapakse (forthcoming) Can Testers Judge Item Difficulty?

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

Pilliner, A E G (1968) "Subjective and Objective Testing" In Davies, A (ed) Language Testing Symposium: A Psycholinguistic Approach. Oxford: Oxford University Press

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