DOCUMENT RESUME

ED 365 151 FL 021 767

AUTHOR	McNamara, T. F.
TITLE	The Role of Item Response Theory in Language Test Validation.
PUB DATE	91
NOTE	21p.; In: Sarinee, Anivan, Ed. Current Developments in Language Testing. Anthology Series 25. Paper presented at the Regional Language Centre Seminar on Language Testing and Language Programme Evaluation (April 9-12, 1990); see FL 021 757.
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Allied Health Occupations; English for Special Purposes; Foreign Countries; *Item Response Theory; *Language Tests; Listening Comprehension; *Second Languages; Testing; *Test Validity; *Vocational English (Second Language)
IDENTIFIERS	Australia; *Occupational English Test; Partial Credit Model

ABSTRACT

The role of item response theory (IRT) in determining the validity of second language tests is examined in the case of one specific test, the listening subtest of the Occupational English Test (OET), used in Australia to measure the language skills of non-native English-speaking health professionals. First, the listening subtest is described. Then the debate over the approp'ateness of IRT use in language testing research is discussed in some detail, with reference to a number of separate studies. Finally, a study of the use of IRT in validating the OET is reported. The study involved analysis, using the Partial Credit model, of data from 196 candidates taking the test in 1987. It investigated whether it is possible to construct a single measurement dimension of listening ability from data from the subtest's two parts, and if the answer is yes, whether the skills tested in the two parts are essentially the same. Results indicate that the test is indeed unidimensional, and support the use of IRT for such analysis. It is also concluded that the kinds of listening tasks in the two subtest parts represent significantly different tasks in terms of level of ability required to deal successfully with them. (MSE)

THE ROLE OF ITEM RESPONSE THEORY IN LANGUAGE TEST VALIDATION

T F McNamara

INTRODUCTION

The last decade has seen increasing use of Item Response theory in the examination of the qualities of language tests. Although it has sometimes been seen exclusively as a tool for improved investigation of the *reliability* of tests (Skehan, 1989), its potential for investigation of aspects of the *validity* of language tests has also been demonstrated (McNamara, 1990). However, the application of IRT in this latter role has in some cases met with objections based on what are claimed to be the unsatisfactory theoretical assumptions of IRT, in particular the so-called 'unidimensionality' assumption (Hamp-Lyons, 1989). In this paper, these issues will be discussed in the context of the analysis of data from an ESP Listening test for health professionals, part of a larger test, the Occupational English Test (OET), recently developed on behalf of the Australian Government (McNamara, 1989b).

The paper is in three sections. First, there is a brief description of the Listening sub-test of the OET. Second, the appropriateness of the use of IRT in language testing research is discussed. Third, the use of IRT in the validation of the Listening sub-test of the OET is reported. In this part of the paper, the issue of unidimensionality is considered in the context of analysis of data from the two parts of this test.

THE LISTENING SUB-TEST OF THE OCCUPATIONAL ENGLISH TEST

The Occupational English Test (McNamara, 1989b) is administered to several hundred immigrant and refugee health professionals wishing to take up practice in Australia each year. The majority of these are medical practitioners, but the following professional groups are also represented: nurses, physiotherapists, occupational therapists, dentists, speech pathologists and veterinary surgeons, among others. Responsibility for administering the test lies with the National Office for Overseas Skills Recognition (NOOSR), part of the Commonwealth

Government's Department of Employment, Education and Training. NOOSR was established in 1989 as an expanded version of what had been until then the Council for Overseas Professional Qualifications (COPQ).

The OET is taken as one of three stages of the process of registration for practice in Australia (the other stages involve pencil-and-paper and practical assessments of relevant clinical knowledge and skills). Prior to 1987, the OET was a test of general English proficiency and was attracting increasing criticism from test takers and test users in terms of its validity and reliability. In response to this, COPQ initiated a series of consultancies on reform of the test. The report on the first of these, which was carried out by a team at Lancaster University, recommended the creation of a test which would (Alderson et al., 1986: 3)

> *assess the ability of candidates to communicate effectively in the workplace.*

A series of further consultancies (McNamara, 1987 ; McNamara, 1988a; McNamara, 1989a) established the form of the new test and developed and trialled materials for it. There are four sub-test, one each for Speaking, Listening, Reading and Writing. The format of the new test is described in McNamara (1989b). The validation of the Speaking and Writing sub-tests is discussed in McNamara (1990).

The Listening sub-test is a 50-minute test in two parts. Part A involves listening to a talk on a professionally relevant subject. There are approximately twelve short answer questions, some with several parts; the maximum score on this part of the test is usually about twenty-five. Part B involves listening to a consultation between a general practitioner and a patient. There are approximately twenty short answer questions (again, some have several parts); the maximum score here is usually twenty-five, giving a total maximum score of approximately fifty on thirty-two items. Because of test security considerations, new materials are developed for each session of the test, which is held twice a year.

Before going on to report on the use of IRT in the validation of the Listening sub-test of the OET, the debate about the appropriateness of the use of IRT in language testing research will be reviewed.

Applications of IRT in language testing

The application of IRT to the area of language testing is relatively recent. Oller (1983) contains no reference to IRT in a wide-ranging collection. By contrast, IRT has featured in a number of studies since the early 1980s. Much of this work has focused on the advantages of IRT over classical theory in investigating the reliability of tests (eg Henning, 1984). More significant is the use of IRT to examine aspects of the validity, in particular the construct validity, of tests.

de Jong and Glas (1987) examined the construct validity of tests of foreign language listening comprehension by comparing the performance of native and non-native speakers on the tests. It was hypothesized in this work that native speakers would have a greater chance of scoring right answers on items: this was largely borne out by the data. Moreover, items identified in the analysis as showing 'misfit' should not show these same properties in relation to native speaker performance as items not showing misfit (that is, on 'misfitting' items native speaker performance will show greater overlap with the performance of non-native speakers); this was also confirmed. The researchers conclude (de Jong and Glas, 1987: 191):

> *The ability to evaluate a given fragment of discourse in order to understand what someone is meaning to say cannot be measured along the same dimension as the ability to understand aurally perceived text at the literal level. Items requiring literal understanding discriminate better between native speakers and non-native learners of a language and are therefore better measures of foreign language listening comprehension.*

This finding is provocative, as it seems to go against current views on the role of inferencing processes and reader/listener schemata is comprehension (cf. Carrell, Devine and Eskey, 1988; Widdowson, 1983; Nunan 1987a). One might argue that the IRT analysis has simply confirmed the erroneous assumption that the essential construct requiring measurement is whatever distinguishes the listening abilities of native- and non-native speakers. An alternative viewpoint is that there will in fact be considerable overlap between the abilities of native- and non-native speakers in higher-level cognitive tasks involved in discourse comprehension. If the analysis of listening test data reveals that all test items fail to lie on a single dimension of listening ability, then this is in itself a valid finding about the multi-dimensional nature of listening comprehension in a foreign language and should not be discounted. The point is that interpretation of the results of IRT analysis must be informed by an *in principle* understanding of the relevant constructs.

In the area of speaking, the use of IRT analysis in the development of the Interview Test of English as a Second Language (ITESL) is reported in Adams, Griffin and Martin, 1987; Griffin, Adams, Martin and Tomlinson, 1988. These authors argue that their research confirms the existence of a hypothesized 'developmental dimension of grammatical competence... in English S[econd] L[anguage] A[cquisition]' (1988: 12). This finding has provoked considerable controversy. Spolsky (1988: 123), in a generally highly favourable review of the test, urges some caution in relation to the claims for its construct validity:

> *The authors use their results to argue for the existence of a grammatical proficiency dimension, but some of the items are somewhat more general. The nouns, verbs and adjectives items for instance are more usually classified as vocabulary. One would have liked to see different kinds of items added until the procedure showed that the limit of the unidimensionality criterion had now been reached.*

Nunan (1988: 56) is quite critical of the test's construct validity, particularly in the light of current research in second language acquisition:

> *The major problem that I have with the test...[is] that it fails adequately to reflect the realities and complexities of language development.*

Elsewhere, Nunan (1987b: 156) is more trenchant:

> *[The test] illustrates quite nicely the dangers of attempting to generate models of second language acquisition by running theoretically unmotivated data from poorly conceptualized tests through a powerful statistical programme.*

Griffin has responded to these criticisms (cf Griffin, 1988 and the discussion in Nunan, 1988). However, more recently, Hamp-Lyons (1989) has added her voice to the criticism of the ITESL. She summarizes her response to the study by Adams, Griffin and Martin (1987) as follows (1989: 117):

> *..This study... is a backward step for both language testing and language teaching.*

She takes the writers to task for failing to characterize properly the dimension of 'grammatical competence' which the study claims to have validated; like Spolsky and Nunan, she finds the inclusion of some content areas puzzling in such a test. She argues against the logic of the design of the research project (1989: 115):

> *Their assumption that if the data fit the psychometric model they de facto validate the model of separable grammatical competence is questionable. If you construct a test to test a single dimension and then find that it does indeed test a single dimension, how can you conclude that this dimension exists independently of other language variables? The unidimensionality, if that is really what it is, is an artifact of the test development.*

On the question of the unidimensionality assumption, Hamp-Lyons (1989: 114) warns the developers of the ITESL test that they have a responsibility to acknowledge

> *...the limitations of the partial credit model, especially the question of the unidimensionality assumption of the partial credit model, the conditions under which that assumption can be said to be violated, and the significance of this for the psycholinguistic questions they are investigating... They need to note that the model is very robust to violations of unidimensionality.*

She further (1989: 116) criticizes the developers of the ITESL for their failure to consider the implications of the results of their test development project for the classroom and the curriculum from which it grew.

Hamp-Lyons's anxieties about the homogeneity of items included in the test, echoed by Nunan and Spolsky, seem well-founded. But this is perhaps simply a question of revision of the test content. More substantially, her point about the responsibilities of test developers to consider the backwash effects of their test instruments is well taken, although some practical uses of the test seem unexceptionable (for example, as part of a placement procedure; cf the discussion reported in McNamara, 1988b: 57-61). Its diagnostic function is perhaps more limited, though again this could probably be improved by revision of the test content (although for a counter view on the feasibility of diagnostic tests of grammar, see Hughes, 1989: 13-14).

However, when Adams, Griffin and Martin (1987: 25) refer to using information derived from the test

> *in monitoring and developing profiles,*

they may be claiming a greater role for the test in the curriculum. If so, this requires justification on a quite different basis, as Hamp-Lyons is right to point out. Again, a *priori* arguments about the proper relationship between testing and teaching must accompany discussion of research findings based on IRT analysis.

A more important issue for this paper is Hamp-Lyons's argument about the unidimensionality assumption. Here it seems that she may have misinterpreted the claims of the model, which *hypothesizes* (but does not assume in the sense of 'take for granted' or 'require') a single dimension of ability and difficulty. Its analysis of test data represents a test of this hypothesis in relation to the data. The function of the fit t-statistics, a feature of IRT analysis, is to indicate the probability of a particular pattern of responses (to an item or on the part of an individual) in the case that this hypothesis is true. Extreme values of t, particularly extreme positive values of t, are an indication that the hypothesis is unlikely to be true for the term or the individual concerned. If items or individuals are found in this way .o be disconfirming the hypothesis, this may be interpreted in a number of ways. In relation to items, it may indicate (1) that the item is poorly constructed; (2) that if the item is well-constructed, it does not form part of the same dimension as defined by other items in the test, and is therefore measuring a different construct or trait. In relation to persons, it may indicate (1) that the performance on a particular item was not indicative of the candidate's ability in general, and may have been the result of irrelevant factors such as fatigue, inattention, failure to take the test item seriously, factors which Henning (1987: 96) groups under the heading of *response validity*; (2) that the ability of the candidates involved cannot be measured appropriately by the test instrument, that the pattern of responses cannot be explained in the same terms as applied to other candidates, that is, there is a heterogeneous test population in terms of the hypothesis under consideration; (3) that there may be surprising gaps in the candidate's knowledge of the areas covered by the test; this information can then be used for diagnostic and remedial purposes.

A further point to note is that the dimension so defined is a *measurement* dimension which is constructed by the analysis, which must be distinguished from the dimensions of underlying knowledge or ability which may be hypothesized on other, theoretical grounds. IRT analyses do not 'discover' or 'reveal' existing underlying dimensions, but rather construct dimensions for the purposes of measurement on the basis of test performance. The relationship between these two conceptions of dimensionality will be discussed further below.

Hamp-Lyons is in effect arguing, then, that IRT analysis is insufficiently sensitive in its ability to detect in the data departures from its hypothesis about an underlying ability-difficulty continuum. The evidence for this claim, she argues, is in a paper by Henning, Hudson and Turner (1985), in which the appropriateness of Rasch analysis with its attempt to construct a single dimension is questioned in the light of the fact that in language test data (Henning, Hudson and Turner, 1985: 142)

> ...examinee performance is confounded with many cognitive and affective test factors such as test wiseness, cognitive style, test-taking strategy, fatigue, motivation and anxiety. Thus, no test can strictly be said to measures one and only one trait.

(In passing, it should be noted that these are not the usual grounds for objection to the supposedly unidimensional nature of performance on language tests, as these factors have been usefully grouped together elsewhere by Henning under the heading of *response validity* (cf above). The more usual argument is that the *linguistic* and *cognitive skills* underlying performance on language tests cannot be conceptualized as being of one type.) Henning et al. examined performance of some three hundred candidates on the UCLA English as a Second Language

Placement Examination. There were 150 multiple choice items, thirty in each of five sub-tests: Listening Comprehension, Reading Comprehension, Grammar Accuracy, Vocabulary Recognition and Writing Error Detection. Relatively few details of each sub-test are provided, although we might conclude that the first two sub-tests focus on language use and the other three on language usage. This assumes that inferencing is required to answer questions in the first two sub-tests; it is of course quite possible that the questions mostly involve processing of literal meaning only, and in that sense to be rather more like the other sub-tests (cf the discussion of this point in relation to de Jong and Glas (1987) above). The data were analysed using the Rasch one-parameter model, and although this is not reported in detail, it is clear from Table two on p. 153 that eleven misfitting items were found, with the distribution over the sub-tests as follows: Listening, 4; Reading, 4; Grammar, 1; Vocabulary, 3; Writing error detection, 3. (Interestingly, the highest numbers of misfitting items were in the Listening and Reading sub-test). One might reasonably conclude that the majority of test items may be used to construct a single continuum of ability and difficulty. We must say 'the majority' because in fact the Rasch analysis does identify a number of items as not contributing to the definition of a single underlying continuum; unfortunately, no analysis is offered of these items, so we are unable to conclude whether they fall into the category of poorly written items or into the category of sound items which define some different kind of ability. It is not clear what this continuum should be called; as stated above, investigation of what is required to answer the items, particularly in the Reading and Listening comprehension sub-test, is needed. In order to gain independent evidence for the Rasch finding of the existence of a single dimension underlying performance on the majority of items in the test, Henning et al. report two other findings. First, factor analytic studies on previous versions of the test showed that the test as a whole demonstrated a single factor solution. Secondly, the application of a technique known as the Bejar technique for exploring the dimensionality of the test battery appeared to confirm the Rasch analysis findings. Subsequently, Henning et al.'s use of the Bejar technique has convincingly been shown to have been unrevealing (Spurling, 1987a; Spurling, 1987b). Henning et al. nevertheless conclude that the fact that a single dimension of ability and difficulty was defined by the Rasch analysis of their data despite the apparent diversity of the language subskills included in the tests shows that Rasch analysis is (Henning, Hudson and Turner, 1985: 152)

> sufficiently robust with regard to the assumption of unidimensionality to permit applications to the development and analysis of language tests.

(Note again in passing that the analysis by this point in the study is examining a rather different aspect of the possible inappropriateness or otherwise of IRT in relation to language test data than that proposed earlier in the study, although now closer to the usual grounds for dispute). The problem here, as Hamp-Lyons is right to point out, is that what Henning et al. call 'robustness' and take to be virtue leads to conclusions which, looked at from another point of view, seem worrying. That is, the unidimensional construct defined by the test analysis seems in some sense to be at odds with the *a priori* construct *validity*, or at least the face validity, of the test being analysed, and at the very least needs further discussion. However, as has been shown above, the results of the IRT analysis in the Henning study are ambiguous, the nature of the tests being analysed is not clear, and the definition of a single construct is plausible on one reading of the sub-tests' content. Clearly, as the results of the de Jong and Glass study show (and whether or not we agree with their interpretation of those results), IRT analysis is capable of defining different dimensions of ability within a test of a single language sub-skill, and is not necessarily 'robust' in that sense at all, that is, the sense that troubles Hamp-Lyons.

In a follow-up study, Henning (1988: 95) found that fit statistics for both items and persons were sensitive to whether they were calculated in unidimensional or multidimensional contexts, that is, they were sensitive to 'violations of unidimensionality'. (In this study, multidimensionality in the data was confirmed by factor analysis.) However, it is not clear why fit statistics should have been used in this study; the measurement model's primary claims are about the estimates of person ability and item difficulty, and it is these estimates which should form the basis of argumentation (cf the advice on this point in relation to item estimates in Wright and Masters, 1982: 114-117).

In fact, the discussions of Hamp-Lyons and Henning are each marked by a failure to distinguish two types of model: a measurement model and a model of the various skills and abilities potentially underlying test performance. These are not at all the same thing. The measurement model posited and tested by IRT analysis deals with the question, 'Does it make sense in measurement terms to sum scores on different parts of the test? Can all items be summed meaningfully? Are all candidates being measured in the same terms?' This is the 'unidimensionality' assumption; the alternative position requires us to say that separate, qualitative statements about performance on each test item, and of each candidate, are the only valid basis for reporting test performance. All tests which involve the summing of scores across different items or different test parts make the same assumption. It should be pointed out, for example, that classical item analysis makes the same 'assumption' of unidimensionality, but lacks tests of this 'assumption' to signal violations of it. As for the interpretation of test scores, this must be done in the light of the our best understanding of the nature of language abilities, that is, in the light of current models of the constructs models such as IRT, and both kinds of analysis have the potential to illuminate the nature of what is being measured in a particular language test.

It seems, then, that Hamp-Lyons's criticisms of IRT on the score of unidimensionality are unwarranted, although, as stated above, results always need to be interpreted in the light of independent theoretical perspective. In fact, independent evidence (of example via factor analysis) may be sought for the conclusions of an IRT analysis when there are grounds for doubting them, for example when they appear to overturn long- or dearly-held beliefs about the nature of aspects of language proficiency. Also, without wishing to enter into Hamp-Lyons (1989: 114) calls

> *the hoary issue of whether language competence is unitary or divisible,*

it is clear that there is likely to be a degree of commonality or shared variance on tests of language proficiency of various types, particularly at advanced levels (cf the discussions in Henning (1989: 98) and de Jong and Henning (1990) of recent evidence in relation to this point).

Hamp-Lyons (1989) contrasts Griffin et al.'s work on the ITESL with a study on writing development by Pollitt and Hutchinson (1987), whose approach she views in a wholly positive light. Analysis of data from performance by children in the middle years of secondary school on a series of writing tasks in English, their mother tongue in most cases, led to the following finding (Pollitt and Hutchinson, 1987: 88):

> *Different writing tasks make different demands, calling on different language functions and setting criteria for competence that are more or less easy to meet.*

Pollitt (in press, quoted in Skehan, 1989: 4)

> *discusses how the scale of difficulty identified by IRT can be related to underlying cognitive stages in the development of a skill.*

For Hamp-Lyons (1989: 113), Pollitt and Hutchinson's work is also significant as an example of a valuable fusion of practical test development and theory building.

Several other studies exist which use the IRT Rating Scale model (Andrich, 1978a; Andrich, 1978b; cf Wright and Masters, 1982) to investigate assessments of writing (Henning and Davidson, 1987; McNamara, 1990), speaking (McNamara, 1990) and student self assessment of a range of language skills (Davidson and Henning, 1985). These will not be considered in detail here, but demonstrate further the potential of IRT to investigate the validity of language assessments.

THE OET LISTENING SUB-TEST: DATA

Data from 196 candidates who took the Listening sub-test in August, 1987 were available for analysis using the Partial Credit Model (Wright and Masters, 1982) with the help of facilities provided by the Australian Council for Education Research. The material used in the test had been trialled and subsequently revised prior to its use in the full session of the OET. Part A of the test consisted of short answer questions on a talk about communication between different groups of health professionals in hospital settings. Part B of the test involved a guided history taking in note form based on a recording of a consultation between a doctor and a patient suffering headaches subsequent to a serious car accident two years previously. Full details of the materials and the trialling of the test can be found in McNamara (in preparation).

The analysis was used to answer the following question:

1. Is it possible to construct a single measurement dimension of 'listening ability' from the data from the test as a whole? Does it make sense to add the scores from the two parts of the Listening sub-test? That is, is the Listening test 'unidimensional'?

2. If the answer to the first question is in the affirmative, can we distinguish the skills involved in the two Parts of the sub-test, or are essentially the same skills involved in both? That is, what does the test tell us about the nature of the listening skills being tapped in the two parts of the sub-test? And from a practical point of view, if both sub-tests measure the same skills, could one part of the sub-test be eliminated in the interests of efficiency?

Two sorts of evidence were available in relation to the first question. Candidates' responses were analysed twice. In the first analysis, data from Parts A and B were combined, and estimates of item difficulty and person ability were calculated. Information about departures from unidimensionality were available in the usual form of information about 'misfitting' items and persons. In the second analysis, Part A and Part B were each treated as separate tests, and estimates of item difficulty and person ability were made on the basis of each test separately. It follows that if the Listening sub-test as a whole is unidimensional, then the estimates of person ability from the two separate Parts should be identical; that is, estimates of person ability should be independent of the part of the test on which that estimate is based. The analysis was carried out using the programme MSTEPS (Wright, Congdon and Rossner, 1987).

Using the data from both parts as a single data set, two candidates who got perfect scores were excluded from the analysis, leaving data from 194 candidates. There were a maximum of forty-nine score points from the thirty-two items. Using data from Part A only, scores from five candidates who got perfect scores or scores of zero were excluded, leaving data from 191 candidates. There were a maximum of twenty-four score points from twelve items. Using data from Part B only, scores of nineteen candidates with perfect scores were excluded, leaving data from 177 candidates. There were a maximum of twenty-five score points from twenty items. Table 1 gives summary statistics from each analysis. The *Test reliability of person separation* (the proportion of the observed variance in logit measurements of ability which is not due to measurement error; Wright and Masters, 1982: 105-106), termed the 'Rasch analogue of the familiar KR20 index' by Pollitt and Hutchinson (1987: 82), is higher for the test as a whole than for either of the two parts treated independently. The figure for the test as a whole is satisfactory (.85).

Table 1 Summary statistics, Listening sub-test

	Parts A and B	Part A	Part B
N	194	191	177
Number of items	32	12	20
Maximum raw score	49	24	25
Mean raw score	34.2	14.4	19.4
S D (raw scores)	9.5	5.3	4.5
Mean logit score	1.46	0.86	1.67
S D (logits)	1.33	1.44	1 25
Mean error (logits)	.48	71	.75
Person separation reliability (like KR-20)	.85	74	.60

Table 2 gives information on misfitting persons and items in each analysis.

Table 2 Numbers of misfitting items and persons, Listening sub-test

	Parts A and B	Part A	Part B
Items	2 (#7, #12)	2 (#7, #12)	1 (#25)
Persons	2	1	5

The analysis reveals that number of misfitting items is low. The same is true for misfitting persons, particularly for the test as a whole and Part A considered independently. Pollitt and Hutchinson (1987: 82) point out that we would normally expect around 2% of candidates to generate fit values above +2.

On this analysis, then, it seems that when the test data are treated as single test, the item and person fit statistics indicate that all the items except two combine to define a single measurement dimension; and the overwhelming majority of candidates can be measured meaningfully in terms of the dimension of ability so constructed. Our first question has been answered in the affirmative.

It follows that if the Listening sub-test as a whole satisfies the unidimensionality assumption, then person ability estimates derived from each of the two parts of the sub-test treated separately should be independent of the Part of the test on which they are made. Two statistical tests were used for this purpose.

The first test was used to investigate the research hypothesis of a perfect correlation between the ability estimates arrived at separately by treating the data from Part A of the test independently of the data from Part B of the test. The correlation between the two sets of ability estimates was calculated, corrected for attenuation by taking into account the observed reliability of the two parts of the test (Part A: .74, Part B: .60 - cf Table 1 above). (The procedure used and its justification are explained in Henning, 1987: 85-86.) Let the ability estimate of Person n on Part A of the test be denoted by bnA and the ability estimate of Person n on Part B of the test be denoted by bnB. The correlation between these two ability estimates, uncorrected for attenuation, was found to be .74. In order to correct for attenuation, we use the formula

$$rxy = \frac{Rxy}{\sqrt{rxx\ ryy}}$$

where rxy = the correlation corrected for attenuation
Rxy = the observed correlation, uncorrected
rxx = the reliability coefficient for the measure of the variable x
ryy = the reliability coefficient for the measure of the variable y

and where if $rxy > 1$, report $rxy = 1$.

The correlation thus corrected for attenuation was found to be > 1, and hence may be reported as 1. This test, then, enables us to reject the hypothesis that there is not a perfect linear relationship between the ability estimates from each part of the test, and thus offers support for the research hypothesis that the true correlation is 1.

The correlation test is only a test of the *linearity* of the relationship between the estimates. As a more rigorous test of the *equality* of the ability estimates, a X^2 test was done. Let the 'true' ability of person n be denoted by ßn. Then b*nA* and b*nB* are estimates of ßn. It follows from maximum likelihood estimation theory (Cramer, 1946) that, because bnA and bnB are maximum likelihood estimators of ßn (in the case when both sets of estimates are centred about a mean of zero),

$$bnA \sim N\ (ßn, e\overset{2}{n}A)$$

where e*nA* is the error of the estimate of the estimate of the ability of Person n on Part A of the test and

$$bnB \sim N\ (ßn1, \overset{2}{e}nB)$$

where e*nB* is the error of the estimate of the ability of Person n on Part B of the test.

From Table 1, the mean logit score on Part B of the test is 1.67, while the mean logit score on Part A of the test is .86. As the mean ability estimates for the scores on each part of the test have thus not been set at zero (due to the fact that items, not people, have been centred), allowance must be made for the relative difficulty of each part of the test (Part B was considerably less difficult than Part A). On average, then, bnB - bnA = .81. It follows that if the hypothesis that the estimates of ability from the two parts of the test are identical is true, then bnB - bnA - .81 = 0. It also follows from above that

$$bnB - bnA - .81 \sim N(0, \overset{2}{e}nB + \overset{2}{e}nA)$$

and thus that

$$\frac{bnB - bnA - .81}{\sqrt{\overset{2}{e}nB + \overset{2}{a}nA}} \sim N\,(0,1)$$

if the differences between the ability estimates (corrected for the relative difficulty of the two parts of the test) are converted to z-scores, as in the above formula. If the hypothesis under consideration is true, then the resulting set of z-scores will have a unit normal distribution; a normal probability plot of these z-scores can be done to confirm the assumption of normality. These z-scores for each candidate are then squared to get a value of X^2 for each candidate. In order to evaluate the hypothesis under consideration for the entire set of scores, then the test statistic is

$$\chi^{2}_{N-1} = \sum_{i=1}^{N} z^{2}$$

where N = 174

The resulting value of X^2 is 155.48,*df* = 173, *p* = .84. (The normal probability plot confirmed that the z-scores were distributed normally). The second statistical test thus enables us to reject the hypothesis that the ability estimates on the two parts of the test are not identical, and thus offers support for the research hypothesis of equality.

The two statistical tests thus provide strong evidence for the assumption of unidimensionality in relation to the test as a whole, and confirm the findings of the analysis of the data from the whole test taken as a single data set. In contrast to the previously mentioned study of Henning (1988), which relied on an analysis of fit statistics, the tests chosen are appropriate, as they depend on ability estimates directly.

Now that the unidimensionality of the test has been confirmed, performance on items on each part of the test may be considered. Figure 1 is a map of the difficulty of items using the data from performance on the test as a whole (N = 194).

Figure 1 Item difficulty map

Difficulty	Item
5.0	
4.0	8
3.0	
2.0	2 3 5 29
	1 12
	11
1.0	25
	15
	7
	16 24
0.0	26
	10 22 23
	6 14 17
	9 18 32
	20 21
-1.0	27 30
	13 28
	4 19
-2.0	
	31
-3.0	

Figure 1 reveals that the two Parts of the test occupy different areas of the map, with some overlap. For example, of the eight most difficult items, seven are from Part A of the test (Part A contains twelve items); conversely, of the eight easiest items, seven are from Part B of the test (Part B has twenty items). It is clear then that differing areas of ability are tapped by the two parts of the test. This is most probably a question of the content of each part; Part A involves following an abstract discourse, whereas Part B involves understanding details of concrete events and personal circumstances in the case history. The two types of listening task can be viewed perhaps in terms of the continua *more or less cognitively demanding and more or less context embedded* proposed by Cummins (1984). The data from the test may be seen as offering support for a similar distinction in the context of listening tasks facing health professionals working through the medium of a second language. The data also offer evidence in support of the content validity of the test, and suggest that the two parts are sufficiently distinct to warrant keeping both. Certainly, in terms of backwash effect, one would not want to remove the part of the test which focuses on the consultation, as face-to face communication with patients is perceived by former test candidates as the most frequent and the most complex of the communication tasks facing them in clinical settings (McNamara, 1989b).

The interpretation offered above is similar in kind to that offered by Pollitt and Hutchinson (1987) of task separation in a test of writing, and further illustrates the potential of IRT for the investigation of issues of validity as well as reliability in language tests (McNamara, 1990).

CONCLUSION

An IRT Partial Credit analysis of a two-part ESP listening test for health professionals has been used in this study to investigate the controversial issue of test unidimensionality, as well as the nature of listening tasks in the test. The analysis involves the use of two independent tests of unidimensionality, and both confirm the finding of the usual analysis of the test data in this case, that is, that it is possible to construct a single dimension using the items on the test for the measurement of listening ability in health professional contexts. This independent confirmation, together with the discussion of the real nature of the issues involved, suggest that the misgivings sometimes voiced about the limitations or indeed the inappropriateness of IRT for the analysis of language test data may not be justified. This is not to suggest, of course, that we should be uncritical of applications of the techniques of IRT analysis.

Moreover, the analysis has shown that the kinds of listening tasks presented to candidates in the two parts of the test represent significantly different tasks in terms of the level of ability required to deal successfully with them. This further confirms the useful role of IRT in the investigation of the content and construct validity of language tests.

REFERENCES

Adams, R J, P E Griffin and L Martin (1987). *A latent trait method for measuring a dimension in second language proficiency. Language Testing 4,1: 9-27*

Alderson, J C, C N Candlin, C M Clapham, D J Martin and C J Weir (1986). *Language proficiency testing for migrant professionals: new directions for the Occupational English Test University of Lancaster.*

Andrich, D (1978a). *A rating formulation for ordered response categories. Psychometrika 43: 561-573.*

Andrich, D (1978b). *Scaling attitude items constructed and scored in the Likert tradition. Educational and Psychological Measurement 38: 665-680.*

Carrell, P L, J Devine and D E Eskey (eds) (1988). *Interactive approaches to second language reading. Cambridge: Cambridge University Press.*

Cramer, H (1946). *Mathematical methods of statistics. Princeton: Princeton University Press.*

Cummins, J (1984). *Wanted: a theoretical framework for relating language proficiency to academic achievement among bilingual studies. In C Rivera (ed.) Language proficiency and academic achievement. Clevedon, Avon: Multilingual Matters, 2-19.*

Davidson, F and G Henning (1985). *A self-rating scale of English difficulty: Rasch scalar analysis of items and rating categories. Language Testing 2,2: 164-179.*

De Jong, J.H.A.L and C.A.W Glas (1987). *Validation of listening comprehension tests using item response theory. Language Testing 4,2: 170-194.*

De Jong, J.H.A.L and G Henning (1990). Testing dimensionality in relation to student proficiency. Paper presented at the Language Testing Research Colloquium, San Francisco, March 2-5.

Griffin P (1988). Tests must be administered as designed: a reply to David Nunan. In T F McNamara (ed.) Language testing colloquium. Selected papers from a Colloquium held at the Horwood Language Centre, University of Melbourne, 24-25 August 1987. Australian Review of Applied Linguistics 11,2: 66-72.

Griffin P E, R J Adams, L Martin and B Tomlinson (1988). An algorithmic approach to prescriptive assessment in English as a Second Language. Language Testing 5,1: 1-18.

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

Henning G (1984). Advantage of latent trait measurement in language testing. Language Testing 1,2: 123-133.

Henning G (1987). A guide to language testing: development, evaluation, research. Cambridge, M A: Newbury House.

Henning G (1988). The influence of test and sample dimensionality on latent trait person ability and item difficulty calibrations. Language Testing 5,1: 8-99.

Henning G (1989). Meanings and implications of the principle of local independence. Language Testing 6,1: 95-108.

Henning G and F Davidson (1987). Scalar analysis of composition ratings. In K M Bailey, T L Dale and R T Clifford (eds) Language testing research. Selected papers from the 1986 colloquium. Monterey, CA: Defense Language Institute, 24-38.

Henning G, T Hudson and J Turner (1985). Item response theory and the assumption of unidimensionality for language tests. Language Testing 2,2: 141-154.

Hughes A (1989). Testing for language teachers. Cambridge: Cambridge University Press.

McNamara T F (1987). *Assessing the language proficiency of health professionals. Recommendations for the reform of the Occupational English Test.* Melbourne: University of Melbourne, Department of Russian and Language Studies.

McNamara T F (1988a). *The development of an English as a Second Language speaking test for health professionals.* Parkville, Victoria: University of Melbourne, Department of Russian and Language Studies.

McNamara T F (ed.) (1988b). *Language testing colloquium. Selected papers from a Colloquium at the Horwood Language Centre, University of Melbourne, 24-25 August 1987. Australian Review of Applied Linguistics 11,2.*

McNamara T F (1989a). *The development of an English as a Second Language test of writing skills for health professionals.* Parkville, Victoria: University of Melbourne, Department of Russian and Language Studies.

McNamara T F (1989b). ESP testing: general and particular. In C N Candlin and T F McNamara (eds) *Language, learning and community.* Sydney, NSW: National Centre for English Language Teaching and Research, Macquarie University, 125-142.

McNamara T F (1990). Item Response Theory and the validation of an ESP test for health professionals. Paper presented at the Language Testing Research Colloquium, San Francisco, March 2-5.

McNamara T F (in preparation). *Assessing the second language proficiency of health professionals.* Ph D thesis, University of Melbourne.

Nunan D (1987a). *Developing discourse comprehension: theory and practice.* Singapore: SEAMEO Regional Langauge Centre.

Nunan D (1987b). Methodological issues in research. In D Nunan (ed.) *Applying second language acquisition research.* Adelaide: National Curriculum Resource Centre, 143-171.

Nunan D (1988). Commentary on the Griffin paper. In T F McNamara (ed.) *Language testing colloquium. Selected papers from a Colloquium held at the Horwood Language Centre, University of Melbourne, 24-25 August 1987. Australian Review of Applied Linguistics 11,2:* 54-65.

Oller J W (ed.) (1983). *Issues in language testing research.* Rowley, M A: Newbury House.

Pollitt A (in press). Diagnostic assessment through item banking. In N Entwhistle (ed.) Handbook of educational ideas and practices. London: Croom Helm.

Pollitt A and C Hutchinson (1987). Calibrated graded assessments: Rasch partial credit analysis of performance in writing. Language Testing 4,1: 72-92.

Skehan P (1989). Language testing part II. Language Teaching 22,1: 1-13.

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.

Spurling S (1987a). Questioning the use of the Bejar method to determine unidimensionality. Language Testing 4,1: 93-95.

Spurling S (1987b). The Bejar Method with an example: a comment on Henning's 'Response to Spurling'. Language Testing 4,2: 221-223.

Widdowson H G (1983). Learning purpose and language use. Oxford: OUP.

Wright B D and G N Masters (1982). Rating scale analysis. Chicago: MESA Press.

Wright B D, R T Congdon and M Rossner (1987). MSTEPS. A Rasch programm for ordered response categories. Chicago, IL: Department of Education, University of Chicago.