DOCUMENT RESUME

ED 339 187 FL 019 389

AUTHOR Griffin, Patrick E.; And Others
TITLE An Alternative Approach to Identifying a Dimension in Second Language Proficiency.
PUB DATE Aug 85
NOTE 22p.; Paper presented at the Annual Meeting of the Applied Linguistics Association of Australia (5th, Queensland, Australia, August 1985).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Elementary Secondary Education; *Evaluation Criteria; Foreign Countries; *Interviews; Item Analysis; *Language Proficiency; Language Tests; Oral Language; *Rating Scales; *Second Languages; Test Construction; *Testing; Test Items
IDENTIFIERS *Partial Credit Model

ABSTRACT

Current practice in language testing has not yet integrated classical test theory with assessment of language skills. In addition, language testing needs to be part of theory development. Lack of sound testing procedures can lead to problems in research design and ultimately, inappropriate theory development. The debate over dimensionality of language and the testing of proficiency illustrates these difficulties. The introduction of confirmatory analysis should improve research on second language learning. In this paper the confirmatory use of a latent trait model (the "partial credit model") is demonstrated as a tool in the development and construct validation of an oral interview test. The model describes the relationship between an individual's proficiency and the difficulty of a language task, allowing for at least two categories of performance, in terms of the probability of a person providing a language sample adequate to earn a given score within a given limit. It was chosen because of its apparent consistency with observations over a wide range of classroom activities. Item analysis and model-to-data fit were conducted on a 29-item interview test given to 270 students. Use of the approach and model was found to be appropriate and valid. A 33-item bibliography is included. (MSE)

AN ALTERNATIVE APPROACH TO IDENTIFYING A DIMENSION IN SECOND LANGUAGE PROFICIENCY

Patrick E. Griffin
Raymond J. Adams
Lyn Martin
and
Barry Thomlinson

VICTORIAN MINISTRY OF EDUCATION
(SCHOOLS DIVISION)

Paper presented at the fifth annual Conference of the Applied Linguistics Association of Australia, University of Queensland, August, 1985.

1. This has been a joint study involving the Curriculum Branch and Adult Migrant Education Services of the Schools Division of the Victorian Ministry of Education.

INTRODUCTION

Approaches to the testing of a second language development have followed teaching methodologies and, in testing as in teaching, there have been swift changes from one methodology to another, with the proponents of each method denouncing the validity of all preceeding methods.

There would appear to be at lest two aspects to many of the problems in language testing. The first is the limitations of classical test theory for the construction and validation of spoken language tests. Current practices suggest that integrated tests which attempt to assess language skills and classical test theory have been unable to provide a technology appropriate to this need. The most authentic and direct of the integrative approaches have generally required an interview format along with a method of judging samples of elicited language according to the degree of accuracy, authenticity and acceptability. In this type of test it is more natural to grade a student's response in a number of categories according to its degree of acceptability in a given situation. But while these approaches have been regarded as the most valid from both a theoretical and a practical perspective, classical test theory which was developed for use with dichotomously scored 'correct'/'incorrect', discrete point test items administered with paper and pencil, has limited applicatons in integrative, authentic language testing. For example, according to Harrison (1983),

> New developments in the theory of language testing since Lado have been slow because the production and statistical justification of multiple choice tests has made other more subjective assessments look weak by comparison. (Harrison, 1983: 84)

A second aspect of the problems, possibly arising from the first, is that tests and other means of measuring second language proficiency are generally used to define variables used in empirical research studies. These studies have aimed at exploring and developing theories of language development, but have been hampered by potential weaknesses in the original measure, based in classical test theory. Stevenson (1985) points out that the difficulties in theory development may arise from a failure to recognise that theory development and testing must work together to further language research. Language testing needs to be part of theory development.

As such, empirical studies depend on variables which have direction, scale and defintite metric properties. When correlational techniques are employed, as in factor and other analyses, the metric requirements are quite stringent. A lack of sound testing procedures can therefore lead to problems in research design and ultimatly to inappropriate theory development. None is more common in Second Language Acquisition Research (SLA) than the debate over dimensionality of language.

DIMENSIONALITY - Exploratory Studies

Numerous definitions and discussions regarding the dimensionality of proficiency and communicative competence exist (e.g., Canale and Swain, 1980; Oller, 1983; Hughes and Porter, 1983; Higgs, 1984; James, 1985; Rivera, 1985) and we do not wish to enter this debate. It remains as an issue that has been at the heart of a range of controversies in second language acquisition (SLA) research. Arguments in the language dimensionality debate range from a denial of any dimensions of proficiency (Pienemann and Johnston, 1985), the hypothesis of a unitary dimension (e.g. Oller, 1983), to a divisible dimension or multi-dimensional models (e.g. Farhady, 1983). Although it has now been widely accepted that neither the unitary nor the divisible dimension hypotheses can be defended in their extreme forms, a comment on the research methodology employed in the debate is worthwhile since the various sides taken appear to be based on statistical and research design bases which might be questionable.

Even the term Proficiency has been challenged in various literature (Johnson & Peineman, 1985). However it is generally accepted as a descriptive term and we have chosen to accept it and adopt it as a developmental description of an individuals relative status in terms of growth of language utility. As many researchers have accepted the use of the term, the debate has focussed on the nature of proficiency development, and whether there is even such a thing as dimensions. Davies (1981) has also made this point.

Johnson and Pieneman's (1985) argument that there are no dimensions of proficiency is difficult to address, as even they report developmental and varitional dimensions. Their developmental dimension is used to illustrate their notion of an implicational relationship. Under these circumstances it is difficult to rationalise zero dimensions in language development or proficiency.

In quantitative approaches an underlying mathematical model of analysis is selected. A collection of measures is then used to demonstrate the validity of that mathematical modelling of language development. In many cases the specificiation of the model is given little or no attention and it is tested with data of unknown measurement properties. It is probably safe to conclude that the underlying mathematical model in the statistical analysis is rarely if ever, considered. If the model or equation underlying the factor analysis ANOVA, regression analysis, or other analyses, were written down and examined as a definition of the way language develops, there would be few who agreed with its appropriateness. However, these analyses are very common. While there is a considerable amount of quality theory generation in the area of language acquisition, the methodology used to test these theories is not of equivalent quality.

In factor analytic studies used to demonstrate dimensionality, it is common for an exploratory principal component analysis or principal factor analysis to be used, with a varimax rotation (e.g. Farhady, 1983). As this procedure is specifically designed to identify multiple factors, which are independent and maximally separated, the discovery of multiple dimensions is not surprising. The indeterminacy of factor analysis virtually assures the researcher that a factor solution will be found. The nature of the solution depends on the technique used to obtain the simple structure.

When measures of a common type are used, it is not surprising that a single dimension is identified. The possibility for this is clearly demonstrated in a multi-trait, multi-method study by Bachman and Palmer (1983). Furthermore, when small case studies involving very few subjects are employed, it is not surprising that no dimensions can be identified. Since the number of cases may only just exceed the number of variables, and hence, small and unstable eigenvalues tend to indicate a lack of specific or general factors. The problem is a statistical one rather than a substantive one, and the nature of identified dimensions in language proficiency development seems to be based on statistical reasoning, which by and large, predetermines the outcome in support of one or another type of theory (Vollmer and Sang, 1983). It is remarkable that the independence of factors, whether single or multiple, is interpreted in dimensional terms. If multiple independent dimensions do exist, then it should be possible to develop teaching programmes around each factor completely isolated and unrelated to programmes for other factors or dimensions. There are not many practitioners who would accept this, but there are numerous research studies which conclude that the independence of factors

is strongly supported by the evidence obtained. The single or multiple factors appear to be manufactured by the analytical methods and the measurements used (Carroll, 1983).

Other non factor-based studies posit theories of language development and proficiency which are based on very small samples (e.g., Schumann, 1975; Krashen, 1977; Pienemann and Johnston, 1985). The argument that five, ten or even twenty cases producing thousands of utterances for analysis, constitute a large data base from which generalisable results are obtainable, is indefensible. There is no doubt that this kind of intensive casework is essential in theory development, but there is a need for more thorough theory testing before external validity can be claimed. Studies based on very small samples tend to yield broad generalisations beyond their external validity.

Problems in research design and the inappropriate use of statistical techniques are not restricted to the language dimensionality debate. For example, Willig (1985) in a review of research in bilingual education commented on the generally poor quality of research in the language area, and the problems that this provides for interpreting the results of many studies.

> The overwhelming message of these findings reflects on the quality of research and evaluation in bilingual education. The unacceptable quality of the major portion of this research is substantiated not only by the information contained in the studies, but also by that not contained in the studies ... Even the kinds of information basic for any reputable research report were frequently missing ... It is imperative that the quality of research and evaluation in bilingual education be upgraded. (Willig, 1985:311)

CONFIRMATORY APPROACHES

The introduction of confirmatory analysis methods into the language area (e.g. Bachman and Palmer, 1983) will play an important role in the improvement of research in the language area, and in particular, theory testing undertaken in SLA research (Stevenson, 1985). It is important however to recognise that the value of this type of analysis will depend on the quality of the underlying measure used in the analysis. It is at the level of constructing these measures that this study is primarily concerned. In this paper the confirmatory use of a latent trait model (the 'partial credit model') is

demonstrated as a tool in the development and construct validation of an oral interview test. It is shown that, following the proposal of a dimension in language proficiency, the use of a dimension-based measurement model such as the 'partial credit model' is appropriate and is capable of beginning the confirmatory. If more than one dimension is thought to exist, each may be defined, constructed and then tested using a dimensional model. These steps should be a priority to any further empirical studies.

SELECTING A DIMENSION

One language development or acquisition model is selected for this study because of its apparent consistency with observations over a wide range of classroom activities, and its ubiquitous nature in literature dealing with language development. This by no means implies that it was taken as a given. The structure of language indeed seemed to be the basis of a division of theory and practice in second language acquisition.

The differences observed in over 60 classroom lessons, the variety of courses, techniques and contexts meant that language acquisition or development models based in achievement of course-specific objectives would not be appropriate for large scale testing. Teacher interviews, analysis of course records, reports and syllabi, coupled with the classroom observations and an extensive review of theoretical literature, suggested that at least one general development area was of general concern. This was the area of language structure or grammar. Details of the identification of this area are given in Griffin, Adams, Martin and Tomlinson (1986). Thus, the study focussed on the development of grammatical skills within various contexts. The aim was then clarified to develop a test of spoken language focussing on the structural elements, while allowing contexts to vary. While many linguists and language instructors may judge this to be a controversial or even an incorrect decision, nevertheless, there are many instances in the literature which theorise that such a dimension exists. The study then sought to define the dimension as an example, without any claim to importance or to dominance among other possible dimensions. Once defined, we have attempted to confirm its existance and demonstrate how such a confirmatory analysis may be used.

In this first investigation we have begun with a dimension that could loosely be termed 'grammatical competence'. This organisation began with a model proposed by Higgs and Clifford (1982). Essentially, the dimension we have attempted to define begins with isolated elements of vocabulary; it then moves onto the use of some basic formulaic language and the basic structures, followed by the more difficult grammatical elements. A complete list of the test objectives and test items can be found in Griffin, et. al. (1986).

A MEASUREMENT FRAMEWORK

It appears that the classical true score and error model of measurement, along with the correlational techniques such as factor analysis, have been unale to deal with measurement problems in language testing and language research. Griffin (1985) proposed the possibility of using a latent trait model, in particular the rating scale model (Andrich, 1978), for use with interview data that are scored in a number of ordered response categories. In this paper we apply the partial credit model (Masters, 1982), a member of the same Rasch family of measurement models (Wright and Masters, 1985), to the design and calibration of oral interview test items. Apart from allowing the analysis of the rating-scale-type data that result from an oral interview, the model provides a framework from which to begin the study of dimensions in language proficiency.

The partial credit model is an extension of the simple Rasch dichotomous model (Rasch, 1960, 1980) that allows for the scoring of items in any number of ordered categories. This is one of its most obvious advantages over classical test theory that normally requires dichotomously scored test items. for example, in the simplest case, a student's response to an interview task or item may be rated 0, 1, 2, according to its degree of increasing acceptability and approriateness. Any number of graded categories may be used but this scale was adopted for this study to illustrate the approach with the simplest multiple category use. Multiple categories allows for varying degrees of correctness rather than the totally correct and incorrect classification allowable with the dichotomous model.

The partial credit model describes the relationship between a person's proficiency (B_u) and the difficulty of a language task (d_{ij}) allowing for at least two categories of performance. This relationship is described in terms of the probability of a person providing a language sample of sufficient adequacy to be given a score (x_i) out of a possible (m) for a particular language task. The model is written as in the fomula below.

$$\pi_{nix} = \frac{\exp \sum_{j=0}^{x_i} (\beta_n - \delta_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} (\beta_n - \delta_{ij})}$$

There are some restrictions placed on the equation to make it mathematically correct, but, in general, the apparently complex equation is inherently simple once the notation is clarified.

The symbol β_n represents the proficiency (or ability) of the person. The symbol δ_{ij} represents the difficulty associated with scoring (i) (in our example j can equal 0, 1 or 2) on item. The nature of the relationship between item difficulty and person ability as specified by the model will become apparent when the item characterisitic curves for specific test items are examined below.

In applying this model it is hypothesised that language tasks based on grammatical structures can be ordered in difficulty from very easy to very difficult. Further, it is assumed that these tasks have their own inherent difficulty, regardless of the student group.

If the student proficiency level is greater than the difficulty of the task then some degree of success (or acceptibility of response) can be expected. Less acceptable responses can be expected if the proficiency level is less than the difficulty of the task. We are hypothesising that a dimension exists, and that it is possible to locate both person and items on that dimension. If we are unable to reject the hypotheses then a comparison of each person to each item will give some confirmation for the dimension defined and yield powerful diagnostics about each student.

By adopting the partial credit model defined in the mathematical equation, we are assuming that students' grammatical competence develops along the dimension, and that we can obtain fine interpretations of responses based on the simple rating scale 0, 1, or 2.

To apply the model it is necessary to define a dimension that is to be measured by the test, construct items to measure on that dimension and then validate the test with the model. Unlike exploratory factor analysis this should not be a 'fishing' exercise. The model allows the hypothesis of the existance of a specified dimension to be explicitly tested in terms of 'goodness of fit'. (See Wright & Masters, 1982) It is not until measurable dimensions of this nature have been defined that it will be possible to examine the dimensionality of language proficiency, such that the debate about dimensionality can be considered as a non-issue until more confirmatory analyses have been completed. A more thorough discussion of these types of mathematical models and their potential for SLA research can be found in Griffin (1985).

TEST DEVELOPMENT

To obtain an adequate data base to undertake these analyses, an interview-based test was developed and administered to 270 students classified as having proficiency levels ranging from 0 to 1+ on the Australian Second Language Proficiency Ratings (Ingram, 1984). A set of 29 items was developed and organised into four subsets of six to eight items each. Each student was then administered one or two of the subsets. A description of the method of item construction and the nature of the items can be found in Griffin, et al (1985, 1986). In brief, an analysis of course content and assessment methods was conducted, based on observations of 60 classroom lessons, together with meetings and teacher workshops at various adult migrant education centres. Course materials were also examined and a series of objectives developed for validation by teachers. The objectives were placed in sequence of instruction, and in estimated order of difficulty for students. Then using Millman's (1974) formulaic approach, a set of expanded objectives were developed and used to generate test items. These were grouped into four subtests, administered to students under standardised interview conditions and scored using the rating scale scoring procedure outlined above.

Results

Each of the item subsets were analyzed using the CREDIT computer porgram[1] (Masters, Wright and Ludlow, 1980) and the results of these item analyses are shown in Table 1. Each item in Table 1 is reported with two difficulty parameters (the columns labelled d(i,1) and d(i,2) respectively), their corresponding standard errors (the columns labelled se(i,1) and se(i,2) respectively) and an index of item fit to the partial credit model. Two difficulties are reported for each item because each item is worth two score points.

TABLE 1

Item Difficulties, Standard Errors and Fit to the Partial Credit Model

Table 1 Item Difficulties, Standard Errors and Fit to the Partial Credit Model

Item	d(i,1)	d(i,2)	se(i,1)	se(i,2)	fit
1.1	-2.1	-1.22	.47	.28	0.41
1.2	-1.06	.41	.30	.26	2.27
1.3	-1.51	-1.63	.43	.30	-0.14
1.4	-0.12	.14	.27	.27	-0.39
1.5	1.09	2.96	.26	.52	0.42
1.6	-0.29	1.44	.26	.30	0.66
1.7	-0.37	0.48	.27	.27	-1.82
1.8	-0.75	2.51	.26	.38	-1.58
N=94					
2.1	-0.86	0.90	.22	.26	0.61
2.2	-1.58	0.36	.25	.22	0.30
2.3	-0.24	0.43	.22	.25	1.69
2.4	-0.93	0.80	.22	.25	-1.10
2.5	-1.49	2.15	.23	.36	-0.87
2.6	-1.02	1.11	.22	.27	-0.81
2.7	-0.92	1.28	.22	.28	-0.25
N=128					
3.1	0.64	0.77	.21	.28	1.67
3.2	-1.58	1.96	.22	.30	-1.28
3.3	-0.39	1.24	.20	.27	-0.44
3.4	-1.48	0.17	.23	.21	1.07
3.5	-2.12	1.47	.25	.26	-0.83
3.6	-0.24	1.20	.20	.27	-1.25
3.7	-2.31	0.03	.28	.20	-0.10
3.8	-1.29	1.93	.21	.30	-0.44
N=147					
4.1	-2.31	0.40	.39	.31	0.95
4.2	-1.58	0.48	.33	.32	1.39
4.4	-1.90	2.12	.33	.46	-2.01
4.5	0.01	0.56	.31	.36	0.43
4.6	1.51	1.66	.37	.60	0.66
4.7	-1.40	0.45	.32	.32	-1.29
N=70					

The difficulty figures range from approximately -3.0 to +3.0 for each subtest. The scale is a logit scale that is, the logarithm of the odds of success at each score level. The scale of measurement is interval in nature and can be transformed to another scale by a simple linear transformation (Wright and Stone, 1980). Most users of test scores prefer to remove negative scores, however we will retain the basic units in our discussion.

The errors of measurement represent the accuracy each score given. Note first that the errors vary for each item. This is a characteristic of the Rasch model, in which traditional global measures of reliability are replaced by specific estimates of error at each point on the dimension. Note also that the values are small compared to the range of difficulty levels and that the errors are smallest in the mid range of the test. That is, each subset of items is most accurate about its midrange which indicates the most appropriate level of administration. In ASLPR terms these would correspond to 0+, 1-, 1 and 1+. Hence the tests give maximum accuracy at levels for whcy they were designed. Further when the overall person scores were correlated with ASLPR, a validity coefficient of 0.67 was obtained. Hence the test measures a dimension strongly related to that assessed by the ASLPR but gives a marked increase in the accuracy of measurement, and move precise diagnostic records.

ITEM CHARACTERISTIC CURVES

The item difficulties reported in Table 1 are best interpreted by reference to the item characteristic curves (ICC's). The ICC's show how the modelled probability of responding in a particular score category varies with ability. Figure 1 shows the item characteristic curve for item 4.4. In this plot there are three curves, one corresponding to each of the three possible scores on the item. The Pr(0) curve shows how the probability of scoring zero is almost one at low levels of ability and then decreases to zero as ability increases. The Pr(2) curve starts at zero for low abilities and then increases continuously and ability increases. The Pr(1) curve increases as ability increases to zero logits and then it decreases as the more able students are most likely to score two. The difficulties -1.90 and 2.12 for this item reported in Table 1 correspond to the intersection of successive probability curves. That is, Pr(0) and Pr(1) intersect at -1.90 and Pr(1) and Pr(2) intersect at 2.12. For abilities less than -1.90 a students most likely score is zero. For -1.90 to 2.12 the students most likely score is one and beyond 2.12 it is two.

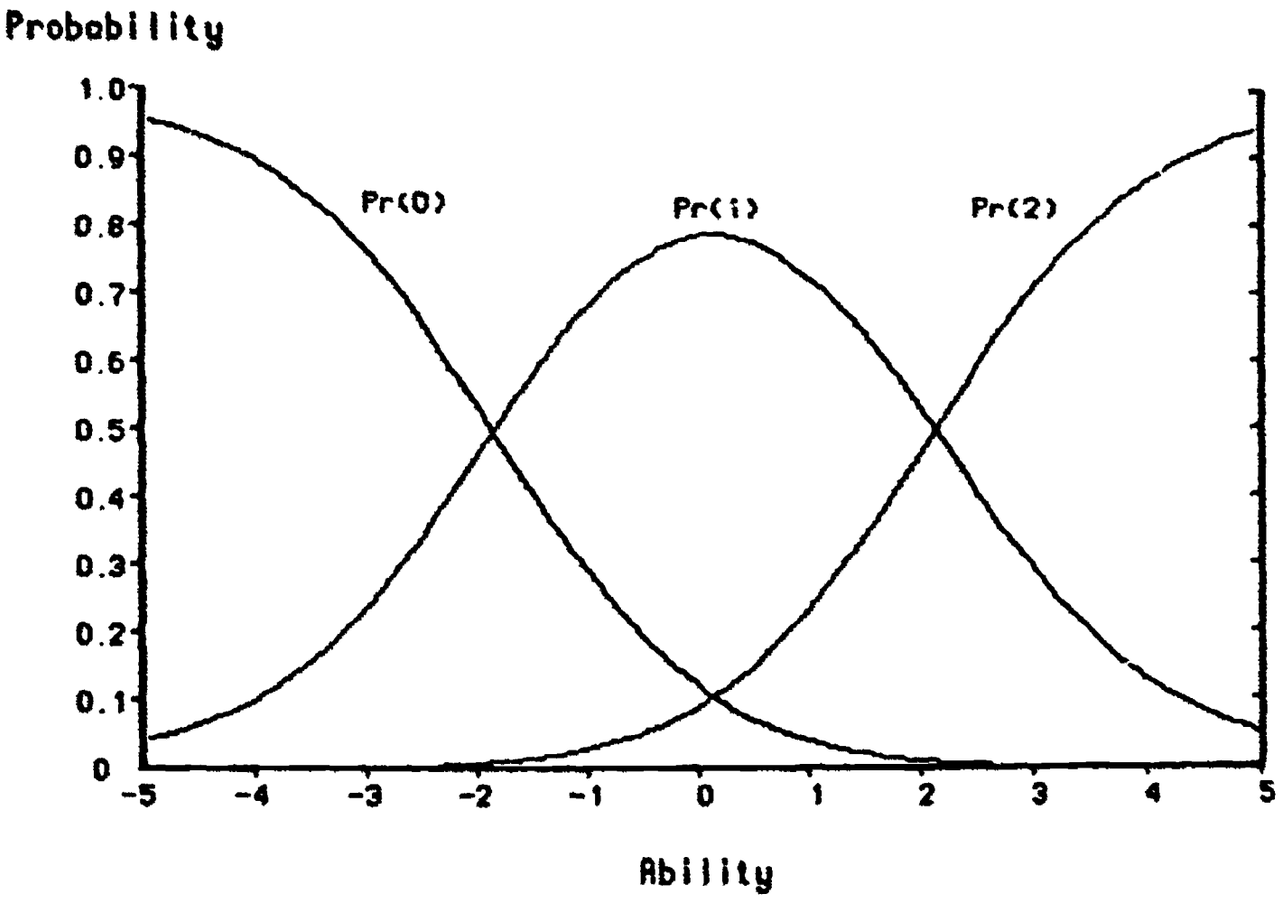

Figure 1 Item characteristic curve for item 4.4

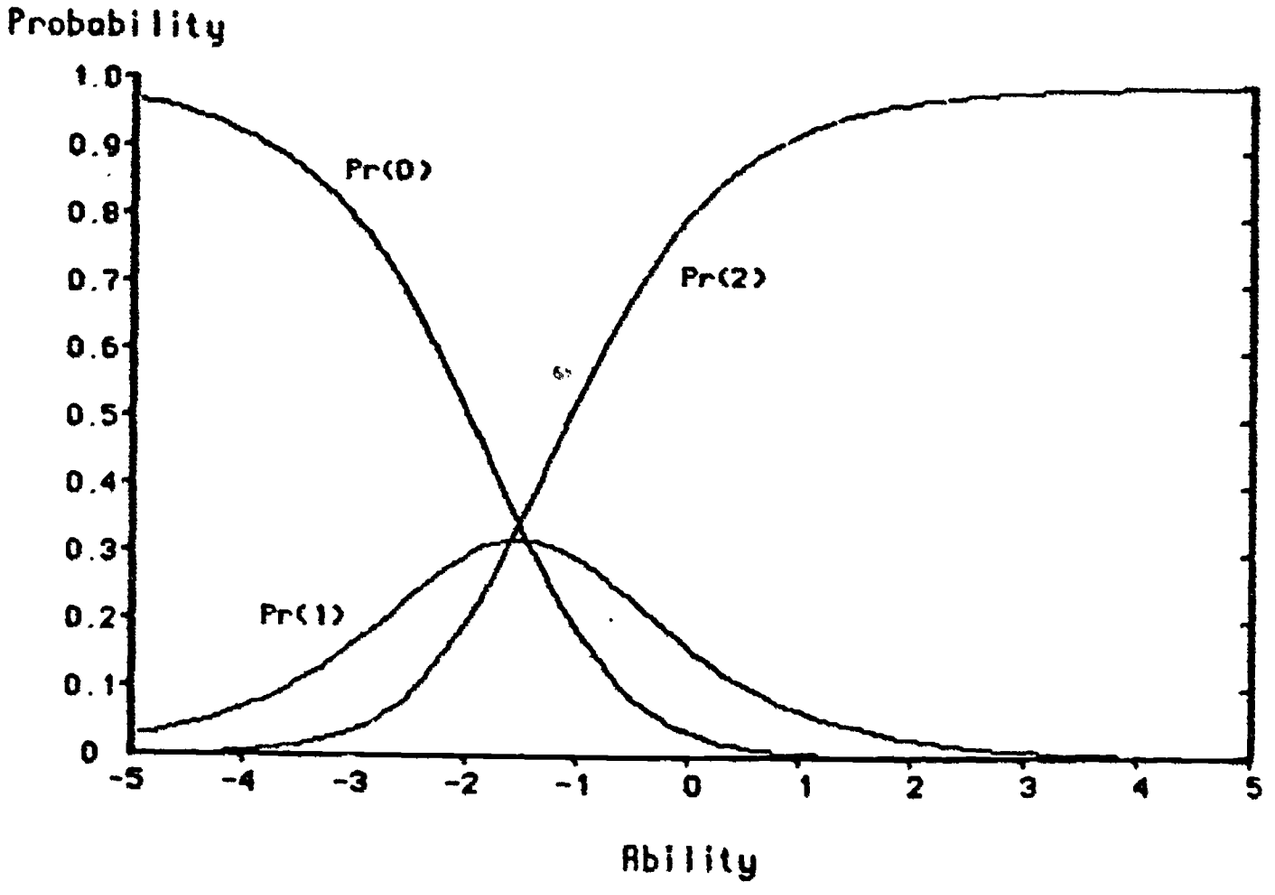

Figure 2 Item Characteristic Curves for Item 1.3

Figure 2 shows the characteristic curves for item 1.3. For this item the difficulties values -1.51 and -1.63 are in reverse order. This means that a score of 1.0 is never the most likely for any student. Students below -1.57 are most likely to score zero and students above -1.57 are most likely to score two.

Since the two difficulties for each item are not independent, the item characteristic curves are essentia' for interpreting the position of the item on the ability dimension. They are also useful for examining the behaviour of items. For example, Griffin et al (1985) used the item characteristics to examine the suitability of scoring criteria for items. If the item characteristic curves are examined in conjunction with scoring criteria it is possible to examine the way different skills are mastered. For eample, item 4.4 has a wide region in which a score of one is most probable. Item 4.4 tests the use of the future tense by using a picture showing a train about to fall from a collapsing bridge. The student is asked, 'What do you think will happen'? The scoring criteria for the item are: clear explanation and consistent use of tense - two points; message clear but no consistency in structure - one point; unintelligible or disjoint words - zero points. In this case the wide region for one score point indicates that many students of varying abilities are able to explain what is happening, but only the best students can use the future tense consistently and appropriately. It would appear that students of relatively low ability can express a sense of futurity but the formal structure is not mastered consistently until fairly high ability levels.

In contrast to item 4.4, item 1.3 (a test of verbs) has no region where one is the most probable response. In this item a student is shown a set of six pictures with people performing various actions. The student is given an appropriate description of the first two pictures and is asked to describe the remaining four pictures. The scoring criteria are: two or more appropriate verbs - two points; one appropriate verb - one point; no appropriate or intelligible verbs - zero points. In this item any form of the verb was considered acceptable. The item characteristic curves show that few students provided only one verb. It would appear that after beginning to understand the use of verbs, students developed a range of verb-based vocabulary almost immediately.

MODEL TO DATA FIT

Since the model is applied to test the hypothesis that a dimension of grammatical competence as defined by the objectives exists, an examination of the model to data fit is crucial. To test the hypothesis each item and each person response pattern is examined to determine whether each fits the dimension assumed by the model. The extent of fit to the model is summarised by a 'fit' statistic. This is discussed in full in Wright and Masters (1982).

The fit statistic for persons and items that conform to the model has an expected value of zero and a standard deviation of about one. The fit of each of the items is shown in the last column of Table 1. When the fit statistic exceeds two, or is less than negative two, there is some doubt about whether this item works in the same way as the other items in the test. Hence we assume that these items do not measure proficiency on the hypothesised dimension. When a large number of items are found to misfit it is unlikely that the set of items is working together to define a measurable dimension, and we would reject the hypothesis of a dimension. Previous research has shown that positive fit (in excess of positive two) usually occurs when an item's score categories do not discriminate between low and high performers as strongly as other items in the test. Negative fit (less than negative two) normally occurs when an item discriminates more highly than other items in the test.

In the analyses reported in Table 1 only items 1.2 and 4.4 were found to have fit statistics outside the range -2.0 to +2.0, hence it appears that each subset is made up of items that are working together. For these circumstances we cannot reject the hypothesis that such structural language tasks can be ordered along a measureable dimension. To investigate possible causes of the misfit of items 1.2 and 4.4, the students' scores on these items were plotted against their proficiency, or ability, as measured by the subset of items which contained the misfitting item. The misfit of item 1.2, plotted in Figure 3 is probably due to the performance of the two students indicated. Both of these students hae scored unity on the item, while according to their overall performance on all the items in the subset, they would be expected to have scored a two. On examination of the recorded interviews for these two students it was found that student 1 was probably scored incorrectly. The item asked the student to describe two persons obviously feeling "hot" and

"cold", in that order. This student was quite confident and quickly responded 'cold and hot' rather than 'hot and cold', then immediately self-corrected. The second student responded 'He is grin' in response to a question were the more appropriate response was 'He is happy'. This vocabulary difficulty was rated down while according to other responses we would expect this student to be able to respond in a fully acceptable manner.

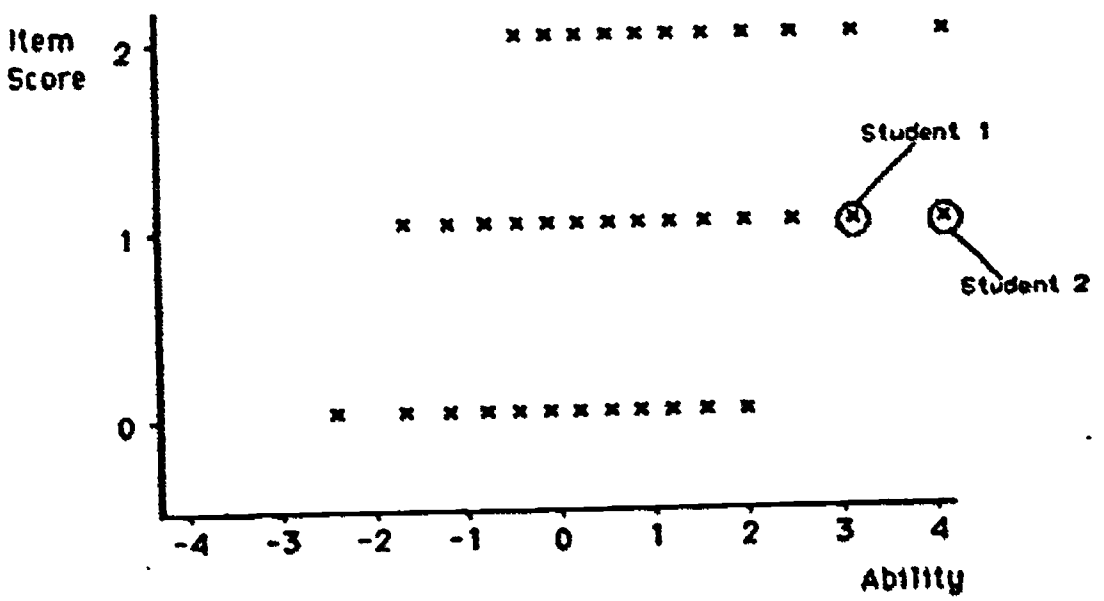

Figure 3 Plot of Score on Item 12 against Ability

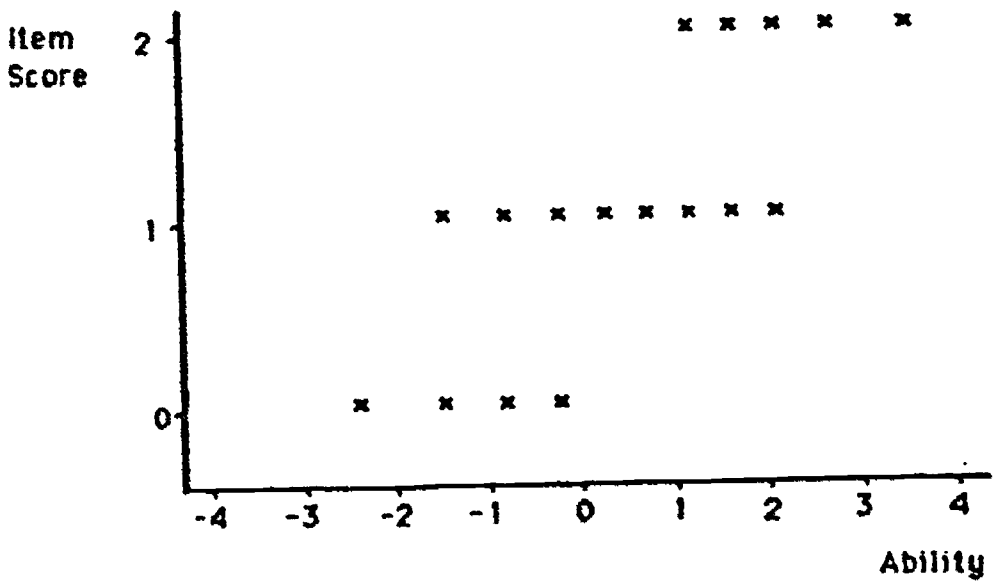

Figure 4 Plot of Score on Item 44 against Ability

Figure 4 shows that item 4.4 has good discrimination. The reason for the misfit of item 4.4 is most likely due to the fact that it discriminates more highly than the other items in the subset. In traditional test analysis procedures this item would probably be regarded as the best but this analysis suggest that its performance should be monitored more closely in future because it is not behaving in the same manner as other items in the test.

Normally, in the case of misfit, items should be deleted from further analysis, however the misfit of these two items is not extreme and does not appear to be due to flaws in the items. They were retained in the subsets for that reason.

In addition to examining the item misfit it is also important to examine any student misfit. As with item fit, the student fit has an expected value of zero and a standard deviation of about unity when the data conform to the model. In Figure 5 below, a bar chart shows the distribution of the student fit statistics. The distribution of fit as shown would appear to support the previous evidence of a strong fit of the model to the data. In particular, note that of the 439 points plotted in Figure 5 only 30 (6.8 percent) lie outside the range -2.0 to 2.0. The response patterns of the 13 students with fit greater than 2.0 do not conform to the model and indicate that for these student the dimension may not be defined in the same way as for the other students ... An intensive follow-up of these students is likely to led to useful diagnostic information about the particular strengths and weaknesses of these students. The fit of less than -2.0 for 17 of the students is due to their response patterns being too orderly and they, as with high negative item fit, are rarely considered as a real problem, but nonetheless, should be monitored.

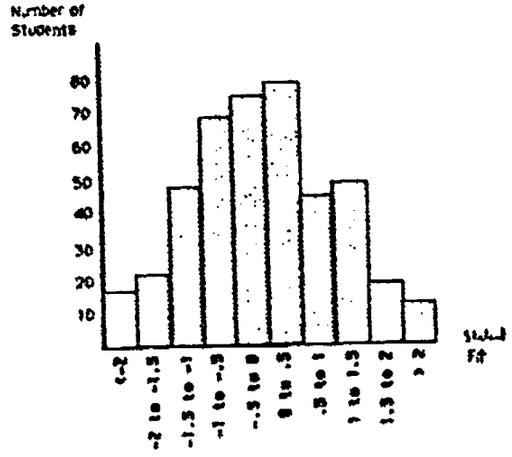

Figure 5 Frequency Distribution of Student Fit to the Model for all Tests

Both the student and item fit statistics support the hypothesis of a single dimension in the data. From a statistical perspective the model has been unable to reject the hypothesis that each of the item subsets can be measured on a single dimension, as defined by the test objectives.

SUMMARY AND CONCLUSIONS

There are several conclusions which might be made from this study. First it is possible to generate tests of spoken language based on generic formulae and which may be scored using a simple rating scale. Second, this rating scale can be defined according to criteria which are robust to variations in scores. This robustness then allows the data to be scaled using the Rasch Rating Scale model. The Latent trait approach has been shown to be successful in defining a grammatical or structural dimension. Further work has been done in demonstrating that the four subjects all measure the same gramattical dimension and this has been reported elsewhere (Griffin, et al 1986). Detailed instructions in administering and interpreting are given elsewhere as well (Griffin 1986).

The third conclusion from the study is that we are able to use the Rasch model as a confirmatory approach to dimensionality. It has not been possible to examine the possibility of constructing a test based on other proposed dimensions or developmental sequences. However, the procedures employed above could and should be applied to the development and validtion of tests hypothesised to test other dimensions, and implicational or developmental sequences.

It is important to note that the identification and confirmation of this dimension is based on first attempting to construct it and the testing whether the constructed variable fits the properties that a measurable dimesnion should have. It does not rely on assumed models and correlational techniques discussed in the introduction of this paper.

Given the success of fitting the model to the data it is possible to argue that a probabilistic approach to the resurement of language development may be adopted for each defensible, definable and demonstrable variable or dimension associated with SLA. This would include Johnson and Peineman's (1985) implicational dimension but it is likely that their Variational dimension is really "lack of fit" (or residuals associated with lack of fit) to their overall implicational dimension.

Given the probabilistic approach of the model it has also been demonstrated that item Characteristic Curves (ICC's), item disnimination, item fit, person fit and other tests of the data each contribute information about language performance. All of this adds to the information about language development and assists in providing profile development and monitoring techniques.

The end result is a measurable dimension of language development or proficiency. Small changes in development may be defined with greater than previously available accuracy. Further, these fine changes in proficiency may be translated into specific skill gains and into the next most likely skill gain.

It is also important to note that in this brief discussion it has not been possible to examine the uses of the test beyond validating the existence of dimensions. Uses of tests, developed with the model, to monitor individual student progress along with the use of fit statistics for individual diagnosis, are quite powerful, and add to the advantages of applying the model to data of this type.

The application of a dimension-based model that can be used with data gathered from oral interviews, clearly has a lot to offer measurement in the language area. If measures are developed from substantive theoretical perspectives and validated via a dimension-based measurement model, they can then be used in confirmatory or exploratory analyses that are powerful tools in the examination of relationships between the dimensions that have been defined and measured.

NOTES

1. The CREDIT program uses the unconditional maximum likelihood procedure described in Wright and Masters (1982) to jointly estimate the student abilities and item difficulties.

REFERENCES

Andrich, D. 1978 A rating formulation for ordered response categories. Psychometrika 43 561-573.

Bachman, L.F. and Palmer, A.S. 1983 The construct validity of the FSI Oral interview. In Oller, J.W. (ed) Issues in Language Testing Research. Rowley. Mass.: Newbury House.

Canale, M. and Swain, M. 1980 Theoretical bases of Communicative Approaches to Second Language Teaching and Testing. Applied Linguistics 11(1), 1-47.

Canale, M. from Communicative Competence to Communicative Language Pedagogy. In J. Richards and R. Schmidt (eds) Language Communication, London; Longvians.

Carroll, J.B. 1983 Psychometric Theory and Language Testing. In Oller, J.W. (ed) Issues in Language Testing Research. Rowley. Mass.: Newbury House.

Davies, A. 1981 Review of Munby, 1978. Tarol Quarterly 15, 3.

Farhady, H. 1983 On the plausibility of the unitary language proficiency factor. In Oller, J.W. (ed) Issues in Language Testing Research. Rowley. Mass.: Newbury House.

Griffin, P.E. 1985 The use of latent trait methods in the calibrario of spoken language in large scale selection-placement programs. In Y. Lee, A. Fok, R. Lord and G. Low (eds) New Directions in Language Testing. London.: Pergamon.

Griffin P.E., Adams R.J., Martin L., Tomlinson B., 1985 The use of latent trait methods to examine second language proficiency. New Horizons, 26, 46-62.

Griffin P.E., Adams R.J., Martin L., Tomlinson B., 1986 Assessing English as a Second Language. Report to the Adult Migrant Education Services. Curriculum Branch, Ministry of Education, Victoria, Melbourne.

Harrison, A. 1983 Communicative testing: jam tomorrow. In A. Hughes and D. porter (eds) Current Developments in Language Testing. London.: Academic Press.

Higgs, T.V. (ed) 1984 Teaching for proficiency, the organising principle. Lincolnwood, Illinois.: National Textbook Company.

Higgs, T.V. and Clifford, R. 1982 The Push Towards Communicative Competence. In T.V. Higgs (ed) Curriculum Competence and the Foreign Language Teacher. Lincolnwood, Illinois.: National Textbook Company.

Hughes, A. and Porter, D. 1983 (eds) Current Developments in Language Testing. London.: Academic Press.

Ingram, D.E. 1984 Australian Second Language Proficiency Ratings. Canberra.: Australian Government Printing Service.

James, C.J. (ed) 1985 Foreign Language Proficiency in the Classroom and Beyond Lincolnwood, Illinois.: National Textbook Company.

Krashen, S. 1977 The Monitor Model for adult second language performance in M. Burt, H. Dulay and M. Finocchiaro (eds) Viewpoints on English as a Second Language. New York.: Regents

Masters, G.N. 1982 A Rasch Model for Partial Credit Scoring. Psychometrika. 47, 149-174

Masters, G.N. 1984 Constructing an item bank using partial credit scoring. Journal of Educational Measurement. 21, 19-32.

Masters, G.N., Wright, B.D. and Ludlow, L., 1980 CREDIT. A Rasch program for ordered response categories. MESA Psychological Laboratory. University of Chicago.

Millman, J. Criterion-referenced measurement. 1974 In W.J. Popham (ed). Evaluation in Education. Berkley, California.: McCutchan.

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

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

Pienemann, M. and Johnston, M. 1985 Factors influencing the development of language proficiency (mimeo).

Rasch, G. 1960 (1980) Probablistic Models for Some Intelligence and Attainment Tests. Copenhagen.: Denmarks Paedegogiske Institute (Chicago, University of Chicago Press).

Rivera, C. (ed) 1984 Communicative competence approaches to language proficiency assessment: Research and Applications. Clevedon.: Multilingual Matters Ltd.

Schumann, J. 1975 Second Language Learning: the pidginization hypothesis. Cited in M. Pienemann and M. Johnston 1985 Factors influencing the development of language proficiency (mimeo).

Stevenson, D.K. 1985 Foreign Language Testing: All of the above. In C.J. James (ed) Practical Applications of Research in Foreign Language Teaching. Lincolnwood Illinois.: National Textbook Company.

Vollmer, H.J. and Sang, F. 1983. Competing hypotheses about secnod lanugage ability: a plea for caution. In Oller, J.W. (ed) Isues in Language Testing Research. Rowley. Mass.: Newbury House.

Willig, A.C. 1985 A Meta-Analysis of Selected Studies on the Effectiveness of Bilingual Education. Review of Educational Research. 55(3), 269-318.

Wright, B.D. and Masters, G.N. 1982 Rating Scale Analysis: Rasch Measurement. Chicago.: MESA Press.

Wright, B.D. and Stone, D. 1980 Best Test Design: Rasch Measurement. Chicago.: MESA Press.

Wright, B.D. and Masters, G.N. 1985 The Essential Process in a family of measurement models. Psychometrika.