DOCUMENT RESUME

ED 342 215 FL 019 492

AUTHOR Griffin, Patrick E.; And Others
TITLE The Development of an Interview Test for Adult Migrants. Proficiency in English as a Second Language.
INSTITUTION Victoria Ministry of Education, West Melbourne (Australia).
REPORT NO ISBN-0-7241-7595-4
PUB DATE 86
NOTE 111p.
PUB TYPE Reports - Descriptive (141)

EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS *English (Second Language); Evaluation Methods; Foreign Countries; *Interviews; Language Proficiency; *Language Tests; *Migrants; Models; *Second Language Learning; Student Evaluation; Student Placement; *Test Construction
IDENTIFIERS Australia; *Interview Test of English as a Second Language

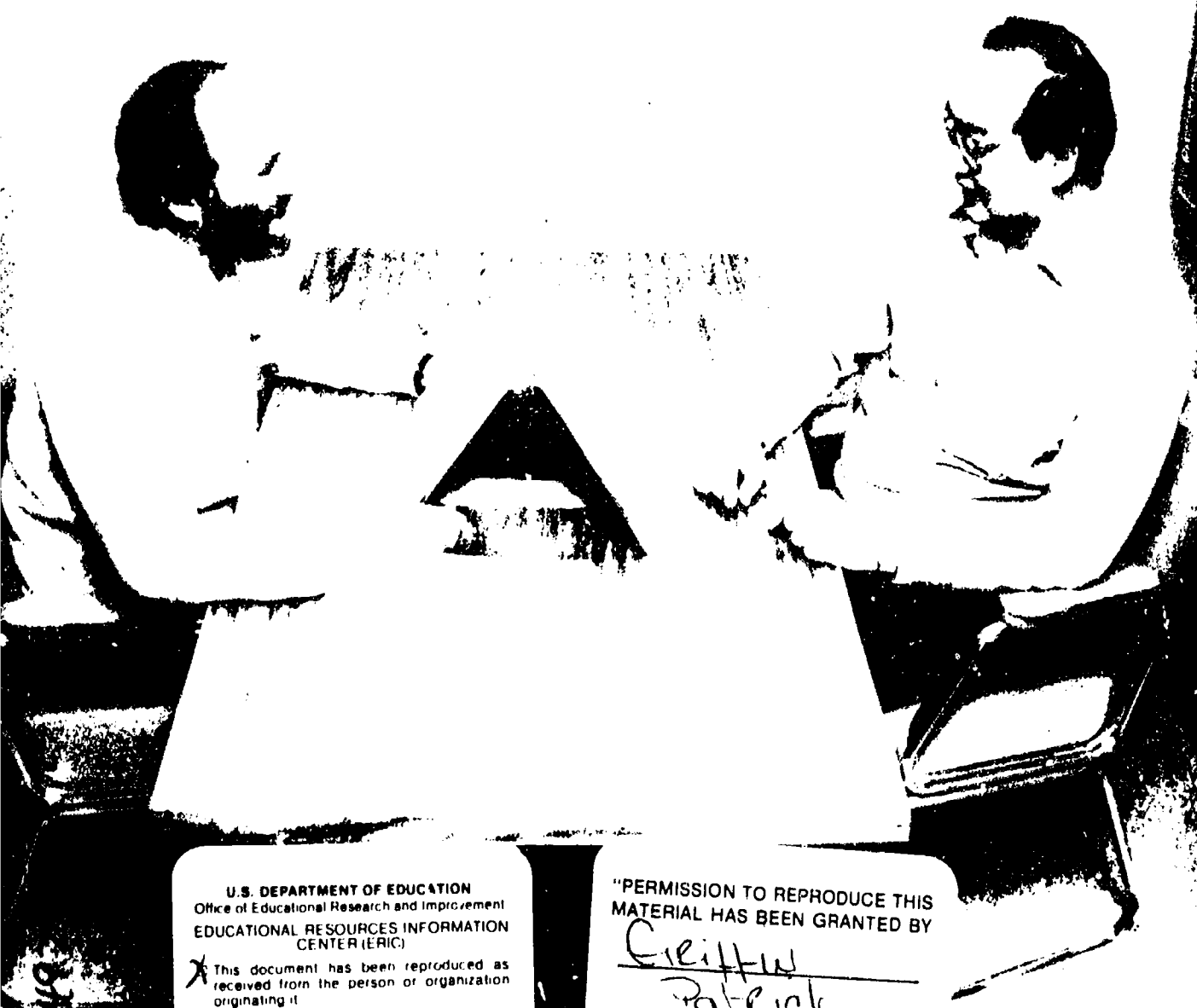
ABSTRACT

The third of three reports resulting from a study of the developing proficiency of adult migrants in English as a Second Language (ESL), this document describes the outcomes of a Victoria, Australia, research and development project to develop mechanisms for implementing evaluation procedures within the Adult Migrant Education Program (AMEP). The primary aims of the project were as follows: (1) to survey two or more education centers to identify testing and assessment tools currently used in ESL instruction; (2) to review program evaluation and student assessment practices; (3) to review the literature on language testing and assessment and then identify the range of components that would be useful in course-specific tests; and (4) to recommend and develop appropriate assessment tools. The first four chapters of this report cover the study background, an overview of program evaluation models, issues in language testing and proficiency, and discussions of measurement models, while the last three chapters cover the organizing principle, construction, and development of a suitable proficiency test. The test developed is called the Interview Test of English as a Second Language (ITESL); it can be used for a variety of purposes; e.g., detailed diagnosis of clients' specific strengths and weaknesses; monitoring the development of clients' oral proficiency; and placement of clients on the basis of oral proficiency. Plans for AMEP evaluation and equivalence tables are appended, and 13 figures supplement the narrative. Contains 95 references. (LB)

PROFICIENCY IN ENGLISH AS A SECOND LANGUAGE

THE DEVELOPMENT OF AN INTERVIEW TEST FOR ADULT MIGRANTS

Ministry of Education (Schools Division), Victoria, 1986



PROFICIENCY IN ENGLISH AS A SECOND LANGUAGE

THE DEVELOPMENT OF AN INTERVIEW TEST FOR ADULT MIGRANTS

MINISTRY OF EDUCATION (SCHOOLS DIVISION), VICTORIA, 1986

THIS REPORT IS ONE OF THE THREE DOCUMENTS PREPARED IN A STUDY OF THE DEVELOPING PROFICIENCY OF ADULT MIGRANTS IN ENGLISH AS A SECOND LANGUAGE.

1. THE DEVELOPMENT OF AN INTERVIEW TEST FOR ADULT MIGRANTS.

2. THE ADMINISTRATION AND GENERATION OF A TEST.

3. AN INTERVIEW TEST OF ENGLISH AS A SECOND LANGUAGE.

Further copies of this publication may be obtained from the Adult Migrant Education Services, Myer House, 250 Elizabeth Street, Melbourne.

National Library of Australia Cataloguing in Publication entry

Proficiency in English as a second language.
The development of an interview test for adult migrants.

Bibliography.
ISBN 0 7241 7595 4.

1. English language - Examinations. 2. English language - Study and teaching - Foreign speakers. I. Griffin, Pat. II. Victoria. Schools Division. Curriculum Branch. III. Victoria. Adult Migrant Education Services. IV. Title: The development of an interview test for adult migrants.

428'.0076

ASCIS Cataloguing in Publication entry (Prepared by SCIS)

Proficiency in English as a second language.
The development of an interview test for adult migrants.

ISBN 0-7241-7595-4.

1. ENGLISH AS A SECOND LANGUAGE. 2. EDUCATIONAL TESTS AND MEASUREMENTS. I. Griffin, Pat. II. Victoria. Schools Division. Curriculum Branch.

428.0076 DDC 19
428 ADDC 11

Without the mastery of the common standard version of a national language, one is inevitably destined to function only at the periphery of national life, and, especially, outside its national and political mainstream. (Gramsci in Tosi, 1984: 167).

FOREWORD

The publication of this report is the culmination of a joint research project involving the Curriculum Branch and Adult Migrant Education Services of the Victorian Ministry of Education.

Few areas of educational measurement have proven as complex as the testing of development in a second language. This study breaks new ground and represents an important application of a particular measurement model to the area of teaching English as a Second Language.

In providing a valuable overview of relevant issues: evaluation models, language proficiency and language testing, the report is a timely and significant contribution to discussion and practice. The testing models which accompany the report make available to teachers practical tools for further application and trialling.

A latent trait model for the analysis of data scored in ordered categories is used as the basis for the construction and analysis of an oral interview test of a dimension of oral proficiency, with important implications for future research.

The recently completed Review of the Adult Migrant Education Program in its committee report, "Towards Active Voice" released in November, 1985, places heavy emphasis on the need to develop systematic planning and evaluation tools for the Program. I am confident that this research has important contributions to make in such a context.

I wish to congratulate Patrick, Lyn, Ray and Barry on their careful and untiring work, as well as the AMES professional staff and their students who have given their valuable time and assistance in refining and trialling materials. This process has been lively and stimulating and I am sure that discussion of the report will be equally stimulating and productive.

Geoff Burke
Supervisor
Adult Migrant Education Services

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

SUMMARY

This report describes the outcomes of a Victorian Ministry of Education research and development project initiated by the Adult Migrant Education Service of Victoria (AMES) and conducted jointly by the Research and Development Section of Curriculum Branch and AMES.

The project began with the recommendations of a working party established to examine methods of evaluation in the AMES. The major goals of the project were to develop mechanisms for the implementation of evaluation procedures across the Adult Migrant Education Program (AMEP). The primary aims of the project were:

1. to survey in two or more education centres to identify testing and assessment tools currently employed in ESL instruction within the AMES;

2. to provide a review of program evaluation and student assessment practices to supplement the first report prepared by the AMES working party;

3. to review the literature in the area of language testing and assessment and identify the range of components which could contribute to course-specific tests as required by teachers;

4. to recommend and develop assessment tools which meet the requirements of the program and to describe these tests in detail.

The survey of current testing practices and tools is presented in chapter 5 of the report. In summary the survey identified only a small amount of suitable material available for testing in English at the proficiency levels most commonly found by the AMES. The Australian Second Language Proficiency Rating (ASLPR) interview, while still widely used for placement, was seen as inappropriate for the finer measurement required to examine improvement and growth in students' language proficiency.

As part of the survey a large number of classes were observed and teachers' reports studied to examine the content of classroom instruction and the assessment techniques employed. In these observations a wide range of teaching styles and methodologies was noted, along with a diversity in the range of content areas covered by different teachers. Due to he understandable emphasis on teaching, teachers' assessment practices were found to be of limited use in the development of interview-based testing materials. As supported by the classroom observation, the teachers' reports identified a number of differing classroom emphases; the one consistent and important item mentioned by teachers was the insistence on the importance of language structure.

A review of appropriate program evaluation models is presented in chapter 2. The method of evaluation proposed by the working party was seen to he closely related to the discrepancy approach described by Provus (1969). The project item rejected the use of a discrepancy approach and throughout the project it stressed that frowth in language proficiency should be emphasised rather than the detailed examination of discrepancies between standards and performance.

In chapters 3, 4 and 5 some literature on language testing and assessment is discussed. As a result of examining the literature, the research team decided to develop interview-based tests of oral proficiency. The dimension underlying the test is based on a model proposed by Higgs and Clifford (1982) and the data collected from classroom observation and the examination of teachers' reports. The example test is in an interview format in which the students are given short oral language tasks and their response is rated according to specified criteria. The Rasch Partial Credit Model was used as the psychometric model for test development. This study is perhaps the first application of this model to oral language tests of this type and has the potential to solve a number of the problems that have existed for the application of sound measurement practices to "authentic language testing".

Chapters 3 and 5 also present the rationale for the adoption of a set of amplified objectives. The objectives were designed so that they could be adapted to the specific requirements of the individual teacher and could therefore be applied to a range of contexts. The objectives were trialled by the research team and they are provided in the accompanying testing manual.

In chapters 6 and 7 details are given regarding the objective development, test onstruction and validation. Along with the associated testing manual they detail how tests can be constructed from the objectives to suit specific courses, and how the test that has been developed can be used to place students in appropriate learning activities, monitor student progress and diagnose individual students' strengths and weaknesses.

The uses and implications of this project are provided in more detail throughout the report. Some of these have been seen as the discrimination of oral proficiency through the application of technological advances made in the application of Rasch models in the area of language proficiency. In this study we have applied the Partial Credit Model, the most general and complex of the Rasch models. The application of this model is only now beginning to be investigated in a range of settings. This is believed to be the first application in the area of language development, and in particular to the speaking skill, which has been considered one of the most difficult measurement areas.

The test developed has been called the Interview Test of English as a Second Language (ITESL). A range of uses for the ITESL has been developed. These include: the detailed diagnosis of clients' specific strengths and weaknesses, the monitoring of development of clients' oral proficiency and the placement of clients on the basis of oral proficiency.

On the basis of an examination of teachers' reports, the observation of teachers' practices and an examination of a large volume of literature, this study has taken a particular stance in the development of test objectives. While the adoption of this stance may be seen as controversial in some areas, the results of the test analyses have clearly supported the theoretical position adopted.

Testing procedures and technology that have been implemented in this study also have implications beyond this project. The methodology discussed and implemented could have important implications for a range of research in the language area. For example, the work of Dulay and Burt (1974a, 1974b) or Pienemann and Johnston (1985) could easily be validiated with the application of a measurement approach similar to that adopted in the study. Furthermore, other dimensions in language proficiency that have been proposed may be tested and validated.

It is hoped that the use of the ITESL will assist the work of the AMEP in solving problems which led to the generation of the project.

ACKNOWLEDGEMENTS

Authors: Patrick E. Griffin, Raymond J. Adams, Lynette Martin; and Barry Tomlinson.

Many teachers and students have assisted through workshops, classroom activities, pilot and trial interviews and the scoring of tests. In particular, the following teachers and AMES staff have made important contributions: Philip McIntyre, Ai Len De Chickera, Patti Wong, Teresa Martin-Lim, Jan Kidman, Sue Hennenberg, Robyn Hughes, Lynda Achren, Lynette Dawson, Chris Corbel, Linda Day, Vivienne Lucena and Grace Waghorn.

Several Migrant Education Centres have contributed by making students available and freeing staff to assist in conducting interviews. These were conducted at the following centres: Myer House, Kuranda, Midway and Collingwood.

We would also like to thank Bill Bell for his contributions to Chapter 1 and Tim McNamara for his valuable input through discussions regarding a number of issues that arose throughout the project.

CHAPTER 1

BACKGROUND TO THE PRESENT STUDY

The Adult Migrant Education Service of Victoria established a working party in 1984 to report on evaluation in the English Language Program (AMES, 1984). The move towards evaluation arose due to toughening attitudes by the Commonwealth Government in its budgeting practices. (Closer scrutiny of resource allocation and use in the Adult Migrant Education Program (AMEP), and a greater degree of accountability was required.) The evaluation of programs required data on learners, and a greater concern for course design and program descriptions.

The Working Party approached the task from the perspective of a Discrepancy Evaluation Model, in which learner performances were to be compared with a specified standard. Four stages of evaluation were defined:

i) Definition of goals and objectives.

ii) Data collection on learner progress and achievement, and identification of discrepancy.

iii) Judgement of identified discrepancy of data to determine where objectives were not met.

iv) Action implementation to redress discrepancy.

The Working Party formulated a style of objective based, in large measure, on the Mager style (1973). Behavioural objectives were examined and judged to be appropriate as a basis for examining change and assessing discrepancy. Formats and procedures were then defined for developing written statements of behavioural objectives and sample course outlines were developed.

Further work was required, however, in translating those objectives into assessment instruments which could provide finer measures of achievement or proficiency than was possible using the currently available techniques.

Since 1979, when early versions were trialled, the AMEP has used the Australian Second Language Proficiency Rating scale (ASLPR) except in NSW. This scale has been used to determine a student's proficiency in the macro skills of listening, speaking, reading and writing. The ASLPR was developed specifically for the AMEP by Ingram (1984), and was based on the scale developed by the United States Foreign Service Institute (School of Language Studies Scale, FSI). The ASLPR scale describes language behaviour at nine proficiency levels along a developmental path from zero to native-like. Each macro skill is defined and described separately. In describing second language development it is also expected that the ASLPR can provide a co-ordinating framework within which program planning and syllabus design can take place (although Ingram warns that it was not specifically designed for that purpose).

In Victoria, the ASLPR has become accepted as an instrument to assist in the measurement of progress in language competence although there has been an identified need to develop an instrument or instruments in a standardised format for finer assessment of development and discrepancies. This need was expressed in the development of the research brief (see Appendix A) and resulted in the commencement of the current project. The focus of the project has been on section A of that brief.

The administration of the AMEP is supported by a number of formal committees, subcommittees, working parties and informal groups at both national and State levels, many of which have a specific curriculum responsibility. At national level, curriculum issues are considered by the National Curriculum Resource Centre (NCRC). This centre, established by the Education Branch, Department of Immigration and Ethnic Affairs, is an independent national unit located within the Adelaide College of Technical and Further Education. The Centre plays a co-ordinating and consultancy role with State AMES centres, provides a materials and syllabus development service, undertakes specific materials and curriculum development projects, provides advice, expertise and support for materials and curriculum development, and disseminates information on current international developments in ESL. It also provides a teacher development service, a publication program, and a national clearing house for materials. While the NCRC assists the development and implementation of curriculum policy in States, it also advises the national AMEP structure, thereby relating particularly to the Joint Commonwealth and State Committee (JCSC) which in turn relates to the national administration of the AMEP, and to the AMEP in Victoria.

From time to time there have been major reviews of the total AMEP program. Until the Galbally Review of Migrant Services and Programs in 1978, the Commonwealth had taken the leading role in determining curriculum rationale. Subsequently, in consultation with the Commonwealth, State AMES have accepted a greater local responsibility. More recently, the entry of other educational providers, particularly TAFE, into the English as a Second Language (ESL) and English for Special Purposes (ESP) fields have broadened the options open to students and have required an extension of the consultative arrangements between them and the AMEP.

In Victoria, consultative arrangements are established between the three AMEP providers - Adult Migrant Education Services, the Royal Melbourne Institute of Technology Language Centre, and the Language Centre at La Trobe University. Separately, the AMEP providers also consult with the State's Child Migrant Education Services and representatives of the TAFE organisation.

Increasingly, at local area levels, representatives of local TAFE Colleges, local AMES Centres, and of local AMES field programs (Home Tutor Scheme, Community Program) meet regularly to plan local delivery arrangements and co-ordinate planning. Within this consultative network can be found, a diversity of approaches to syllabus design, methodology and philosophic perceptions of ESL, generally reflecting the volatility of recent academic developments in ESL.

Most recently there have been moves towards a new definition of curriculum for the AMEP, particularly through the work of the National Curriculum Resource Centre and through the Professional Development Subcommittee of the Joint States and Commonwealth Committee. However, at this stage, there is not yet an overall generally accepted and implemented curriculum rationale in the AMEP. There is a generally held commitment to "student needs" approaches, and within Victoria, where TAFE and AMES are quite separate organisations, an understanding that, where possible, AMES accepts responsibility for lower-level learners. TAFE provides service to higher-level students and to those requiring ESP. There is, however, some indication that this may change in the near future.

Over the years, the AMEP has broadened the range of its programs, which now include, in addition to courses conducted in AMES Centres, field programs, distance learning and self-access options.

The field programs, the Home Tutor Scheme, the English and the Workplace Program, the joint AMES and Commonwealth Employment Service "Jobseekers" programs, and the local suburban and country community classes attract specific groups who are unable to attend major centres. At least initial entry into classes in these programs is further dependent on ASLPR homogeneity (classes being conducted at different ASLPR levels), or common first language, or ethnicity according to opportunity and need, in the community programs.

Major AMES Centres and the Community Program offer "on arrival" courses for new arrivals. The syllabus combines language and information relevant to newcomers.

The major AMES Centres also offer a range of general courses at graded ASLPR levels which allow progression. Courses which focus on literacy, pronunciation, grammar, and general oracy skills are conducted in these centres.

Increasingly, all local providers - AMES, TAFE and CMES in particular - are co-ordinating their programming to allow improved progression and choice by students. A common comprehensive referral system is used in the counselling of students by all providers. Student profiles/histories including ASLPR assessments, are maintained in the (national) AMEP computerised information system, which can be used for selection of homogeneous class groups according to ASLPR and a range of other criteria - purpose, age, sex, first language and ethnicity for example.

Placement and Assessment of Students

The student, the registrar, the teacher in charge (organiser or principal), and the teacher all play a part in considering which course/class is the most appropriate; although, naturally, it is for the student to finally accept or reject any recommendation offered.

The student may apply for admission to an advertised specific course for which the published entry criteria make him eligible - self-selection.

The principal, with a knowledge of the locality and of students' needs has a degree of autonomy, within agreed guidelines of program budgeting constraints, in planning local courses for local students.

The registrar, with the aid of the ASLPR assessment for each student and with access to the referral system and student profiles, can counsel students in their best interests.

The teacher, making a final selection for a particular class group, will counsel or refer other applicants to other classes. The information system presently containing student profiles, including ASLPR assessment data, will also include referral system information in the near future. The referral system is presently manual, with access provided throughout the AMEP and TAFE organisations.

Currently students who enter the AMEP may remain in courses until they have reached level 3 on the ASLPR scale, and are counselled appropriately. In practice, because of the voluntary attendance of students in the AMEP, many rarely remain for long continuous periods. A great deal of attention is currently being given to improving the sequential progression of courses and to extending their length, in order to encourage longer continuous learning periods. However, a substantial part of the total AMEP will probably remain flexible, with non-sequential courses, in order to accommodate the irregular patterns of withdrawal and return which characterise an adult student body on whom a great number of external factors - employment and family care in particular - have their effect. It is extremely rare that any student eligible for tuition and enrolled in the AMEP has been required to leave.

It is unlikely that a student will be referred outside the AMEP until level 3 on the ASLPR is achieved. Students who are assessed as requiring additional instruction after the AMEP courses are referred to:

1. TAFE, if the ASLPR assessments average at 3 or above.

2. TAFE, for ESP programs conducted by them.

3. TAFE community classes where the rationalisation of delivery has resulted in TAFE rather than AMES conducting classes in a given area.

Against this background, the Adult Migrant Education Service in Victoria approached the Curriculum Branch and proposed a joint project with a dual purpose. The first aim was to identify strategies for the assessment and placement of clients in appropriate lessons, courses and programs within the Adult Migrant Education Service and the second was to evaluate course ıtlines. The original Research Brief is appended.

The goal of the proposed research project was:

> "To undertake and report on a trial implementation of an evaluation model outlined in the paper 'Evaluation in the Adult Migrant Education Program (Stage 1)' and to develop mechanisms and testing instruments necessary for implementation across the AMEP". (See Appendix A)

Expected Outcomes

As detailed in Appendix A, the project was expected to achieve four outcomes:

1. to survey practices in two or more education centres to identify testing and assessment tools currently employed within AMES. This is described in Chapter 5.

2. to provide a review of program evaluation and student assessment practices to supplement the first report prepared by the AMES Working Party. This is presented in Chapter 2.

3. to review the literature in the area of testing and assessment and identify a range of components which could contribute to course-specific tests as required by teachers. This is reported in Chapters 3 and 5, together with the rationale for a decision to opt for amplified objectives in the spoken language. These objectives have been trialled by the team in developing a proficiency measure and this and the accompanying reports detail procedures for developing course-specific tests.

4. to recommend the development of assessment tools which meet the requirements of the program and to describe these tests in detail. This has become the major focus of this study and the accompanying documents present a sample test and its administration manual.

CHAPTER 2

PROGRAM EVALUATION MODELS: AN OVERVIEW

Curriculum evaluation is taken to mean the collection of information on the curriculum for the purpose of decision making to improve the curriculum. The focus of the evaluation can be on clients, on teaching or on centres; the techniques can include measurement assessment or observation or case-study. What distinguishes curriculum evaluation from other evaluations with the same focus or the same technique is the purpose of its use: to improve the quality of teaching and learning.

Evaluation information can be used for a variety of purposes:

- to give information to students on their progress;
- to give information to teachers on their effectiveness;
- to diagnose individual student strengths and weaknesses;
- to select students for particular teaching or administrative purposes;
- to provide information on achievement levels for internal or external audiences.

There are many other purposes but these concentrate on improvement of curriculum and on student learning.

In order to examine program evaluation more closely, several evaluation models were reviewed, emphasising their relevance to curriculum improvements.

Contemporary Evaluation Models

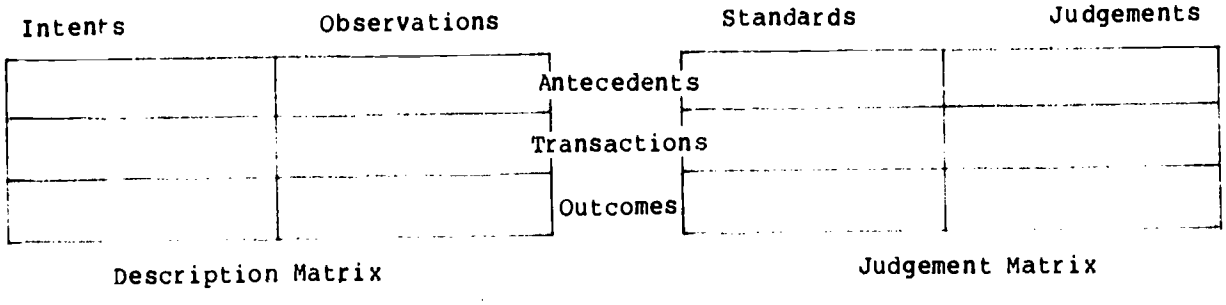
The models to be discussed in this section were those proposed by Stake (1967), Scriven (1967), Provus (1969), Hammond (undated), Stufflebeam (1968), Tyler (1958), Alkin (1969), Parlett and Hamilton (1976), and the Professional Judgement Model exemplified by school accreditation programs. Each model will be discussed briefly; no attempt will be made to detail all the unique features or concepts included in each model. However, the discussion that follows should pinpoint or indicate important differences between the models

and the major evaluation activities suggested by each. It should also provide sufficient information about each model so that its application in each evaluation problem posed in the simulation materials can be identified.

Stake's Model

Stake's model was first proposed in 1967. It is focused on the description and judgement of ongoing educational programs. The evaluator is required to collect, process, and report descriptive and judgemental data about the program setting or expectations for it (e.g. teachers, subject matter specialists, parents, students).

The two types of information -- descriptive and judgemental -- are used by Stake to produce two data collection matrices. These matrices are shown diagrammatically in the figure below. The description matrix is divided into two classes of information -- intents and observations. Intents are goals or objectives stated in any form amenable to evaluation. Observations are what the evaluator learns through direct observation, unobtrusive measures or administration of specific data collection instruments. The judgement matrix is also divided into two classes of information -- standards and judgements. Standards refer to either absolute or relative external standards (criteria) which might be used to judge the worth of whatever is being evaluated. Judgements include deciding whether relative or absolute standards should be applied, assigning weights to various standards, and judging the merit of the program or product under consideration.



Stake's Model

Within each matrix, there are three types of information specified: antecedents, transactions, and outcomes.

Antecedents are those conditions that existed prior to program implementation and which are likely to relate to the outcomes (e.g. student abilities, facilities).

Transactions refer to all the processes that occur during the implementation of the program (e.g. student-teacher interactions, teacher hostility towards a new innovation) etc.

Outcomes refer simply to all consequences of the program. These may be planned or unplanned.

Two other concepts central to Stake's model are contingency and congruence. Contingencies are little more than "if-then" relationships, based on data or logic, used to relate antecedents, transactions and outcomes. Evaluators might look for contingencies between those three types of information by posing such questions such as, "Given this set of conditions (antecedents), and this set of activities and events (transactions), what would you expect to happen (outcomes)?" If one is assessing contingencies between intended program elements, the contingencies are logical contingencies. If, however, one is assessing contingencies for observed elements of the program, they are based on data and are empirical contingencies.

At the same time as the evaluator is assessing contingencies, he must also assess the congruence between intents and observations and between standards and judgements. This simply refers to the identification of discrepancies which exist between intents and observations. As such, the Stake model would be appropriate for an AMES evaluation given the emphasis on discrepancy.

In summary, the major emphasis of Stake's model is the description of describing intents and observations on program antecedents, transactions, and outcomes and the judgement of these against absolute and/or relative standards, in order to assess the merit of the program.

Stufflebeam's (CIPP) Model

Stufflebeam's original model first appeared in 1968. In this model, evaluation is aimed specifically at providing information to serve the decision-making process. In Stufflebeam's view, decision making cannot be rational unless the decision maker can (a) identify the alternatives available in making each decision, and (b) assess the relative merit of each alternative in relation to specific criteria. The role of the evaluator is to collect and supply appropriate information about all available alternatives to enable the decision maker to make sound judgements among them.

Stufflebeam sees decisions as falling into four major classes -- planning, programming, implementing and recycling decisions. Planning decisions are those related to the specification of the domain and setting of major goals and specific objectives for the program. Programming (structuring) decisions are those related to the actual, ongoing conduct of the program. Implementing decisions are those related to directing programmed activities. Recycling decisions are decisions made at the end of a full program cycle about whether to terminate, continue, or modify the program.

For each class of decisions, Stufflebeam proposes a parallel type of evaluation: Context, Input, Process and Product (CIPP), respecti ely.

Context evaluation consists of those activities which define the operational context or system, identify intended outcomes, measure or observe actual outcomes, compare intended and actual outcomes to identify discrepancies (needs), postulate problems underlying identified needs, and establish objectives which, if attained, would solve the problems and thus satisfy the needs.

Input evaluation includes identifying and assessing alternative strategies and designs for attaining program objectives, with specific focus on system capabilities, cost benefits, and potential barriers to success in relation to each alternative.

Process evaluation is aimed at monitoring the ongoing program to detect deviations from the program design, to watch for predicted barriers to success, and to remain alert to unanticipated problems that arise. Immediate feedback to program operators is an essential feature of this type of evaluation.

Product evaluation is terminal evaluation aimed at assessing, on the basis of specified criteria, the extent to which program objectives have been met.

In all stages of evaluation, Stufflebeam sees the evaluator and the decision maker working closely together to assure relevance of the evaluation to the decision maker's needs. For each type of evaluation, a series of steps for designing an evaluation is proposed as follows: focusing the evaluation, collecting the information, organising the information, analysing the information, reporting the information, and administering the evaluation.

In summary, the Stufflebeam CIPP model is aimed at delineating, collecting and reporting information to help the decision maker make intelligent judgements about decision alternatives faced. To tnat end, context, input, process and product evaluation are proposed to provide data in relation to planning, programming, implementing, and recycling decisions.

Provus's (Discrepancy) Model

Perhaps the best presentation of Provus's model is that included in the 1969 yearbook of the Naticnal Society for the Study of Education. The rationale of this model is similar to the CIPP model in that it focuses heavily on providing information to support decision making. The model is designed primarily for programs already staffed and underway. In such programs, Provus sees evaluation occurring at four major stages: definition, installation, process and product. A fifth stage, cost-benefit analysis, is seen as optional to the evaluator who has completed the first four stages. In the first stage, definition, the basic concern is in defining or delineating the precise program or program components to be evaluated. In the second stage, installation, the concern is whether or not the program is installed in accordance with its basic definition. In the process stage, the crucial question is whether or not the enabling objectives are being met. In the output stage the model focuses on the costs of the program in relation to the benefits received.

In each of the first four stages of evaluation, standards are compared with program performance to produce discrepancy information. If discrepancies exist, changes are made in either the program performance or the standards. This discrepancy information is essential to decisions about proceeding to the next evaluation stage, recycling or terminating the program. The end result of Stage I, the program definition, becomes the standard for Stage II, installation, and so on. Provus (1969, p.247) indicates that evaluation "...consists of moving through stages and content categories in such a way as to facilitate a comparison of program performances with standards while at the same time identifying standards to be used for future comparisons."

The three major content categories required in this model are input, process, and output, which parallel closely the antecedents, transactions and outcomes proposed by Stake. For each content category, two pervasive elements that must be examined are time and cost.

The role of the evaluator in Provus's view is that of a team member who works with the administrator and program staff to use evaluation for program improvement. That is, the process of evaluation is performed not only by the evaluator, but the evaluator in co-operation with staff in the program unit.

In summary, the Provus model requires comparison of standards and program performance so as to provide discrepancy information at each of four stages of evaluation, definition, installation, process and output, and for each of three major content categories, input, process and output. Identified discrepancies result in changes in either the standard or program performance so as to eliminate the discrepancy before proceeding to the next stage of evaluation.

Scriven's Model

It may be a misnomer to refer to Scriven's work as an "evaluation model", despite the fact that his 1967 paper in the AERA monograph series on curriculum evaluation (Scriven, 1967) has proved to be one of the seminal works in the field. Instead, it might be best viewed as a collection of insights about evaluation that have great utility for evaluation personnel.

Of the many concepts proposed by Scriven, three should perhaps be stressed as most relevant here. The first is the distinction between formative and summative evaluation. The second is the emphasis on judgement as an essential part of the evaluator's role. The third is the proposition that the worth of goals or objectives must also be judged by the evaluator. Each of these ideas is discussed briefly below.

Scriven differentiates between two basic types of evaluation, which differ according to the audiences for whom the report is intended. Formative evaluation is evaluation aimed at assessing the quality of an educational product or practice during its development, with the producer being the primary audience. Such evaluation is viewed by Scriven as an appropriate role for an internal evaluator -- a person employed by the producer. Summative evaluation is terminal evaluation aimed at judging the merit of the completely developed product or practice, with the consumer being the primary audience for the evaluative information. The summative evaluator role is best played by a person outside the producing agency, since to do otherwise would lessen the credibility of the evaluation.

A second idea emphasised by Scriven is that an essential part of the role of the evaluator is to make judgements about the merit of the entity he is evaluating. Quite unlike Stufflebeam and Alkin, Scriven feels that the evaluator abdicates a portion of his responsibility if he collects and reports evaluative information to the decision maker without also including his honest appraisal of the worth of whatever is being evaluated. In short, judgement, for Scriven, is the sine qua non of the evaluator's role.

The position taken by Scriven is that it is not sufficient to merely assess whether or not the goals or objectives of the program have been met; it is also essential that the evaluator evaluates the worth of the goals or objectives themselves. In discussing extreme relativism in evaluation, Scriven writes:

> The slogan became: "How well does the cause achieve its goals?" instead of "How good is the cause?" But it is obvious that if the goals aren't worth achieving then it is uninteresting to see how well they are achieved... Thus evaluation proper must include, as an equal partner with the measuring of performance against goals, procedures for the evaluation of the goals. (Scriven, 1967 : 51-52)

Although not proposed as a formal model, the set of constraints discussed by Scriven doubtlessly has had as much impact on the field of evaluation as any of the models discussed in this section.

Tyler's Model

Tyler's model for evaluation of learning experiences, one of the earliest evaluation models, was originally developed during the evaluation of the Eight-year Study in the 1930s and early 1940s.

In Tyler's model, evaluation is proposed as an adjunct to the curriculum development process. In fact, the evaluator is the curriculum specialist. The basic rationale for the model is that curriculum (learning experiences, in Tyler's terminology) should be evaluated by comparing student performance with clearly specified behavioural goals established for the curriculum. The basic model consists of six major steps, each of which is discussed below.

The first step in Tyler's model consists of of establishing the broad goals or objectives of the program. The bases for setting goals and objectives are knowledge about pupil entry behaviours, analysis of societal trends and expectations, the nature of knowledge in relevant fields of study, theories of learning and instruction, and the educational philosophy of the school.

Once general objectives have been established, the second step is to classify objectives into a taxonomy aimed at achieving an economy of thought and action.

The third step in Tyler's model is that of defining objectives in behavioural terms. This step in the model has had more long-range impact on evaluators and curriculum specialists than any other aspect of the model. The dependence of several contemporary evaluation models on specific, behavioural statements of objectives, as well as the often misunderstood practice of expressing everything in "behavioural" terms, stems directly from the emphasis in this influential model. It is implicit in Tyler's model that instructional objectives must be pupil-orientated -- i.e., they should contain references not only to course content but also to mental processes to be applied by pupils.

The fourth step is to suggest situations in which achievement of the objectives can be shown. The fifth is to develop or select measurement techniques (questionnaires, rating scales, standardised tests, etc.).

The sixth step in the model embodies the essence of evaluation as Tyler sees it. In this stage, student performance data are gathered and compared with the behaviourally stated objectives. In this fashion, decisions are made about whether or not the curriculum is achieving the desired outcomes. Presumably, appropriate modifications would be made on the basis of such information.

In summary, Tyler's model depends on comparing student performance with specific, behaviourally stated objectives to ascertain the effectiveness of the learning experiences provided by the program. This model has been extremely influential on subsequent writing about evaluation, including the development of concepts in some of the models discussed earlier in this section.

Tyler's report is interesting in two senses: first, it establishes a pattern for evaluation which is still very influential - the pattern which specifies statements of purpose, clarified in a variety of ways, statements of teaching methods, statements of pupil activities and statements of outcomes which are used as part of the evaluation theme - and, secondly, a wide variety of records are used in an attempt to give a complete picture. Tyler uses pupil performance, sociograms, interviews, pupil diaries, case studies -- almost all of the sorts of information which we would currently suggest as relevant, were in one way or another in his approach. This is not typical of the forms of evaluation used at that time -- it is not even typical of the way people describe Tyler today because he is often used as an arch-example of the limited evaluator, which does not reflect the way he actually operated. An important extension of Tyler's work was by Bloom (1956) and others in their attempt to provide means of specifying objectives in a more precise and consistent way and of indicating levels of complexity in a way that was communicable between different groups.

Quite early in the American scene, emanating particularly from the work of Hastings (1969) and Cronbach (1963), was the view that the psychometric model was only one way of looking at and evaluating situations and particularly people's operations. This view suggested that sociology, history and anthropology offer a number of other relevant patterns. This concept was developed further by Robert Stake (1967), who stressed both the formal and informal aspects of evaluation and introduced the concept of evaluation as portrayal.

The book by Wiseman and Pidgeon (1970) builds quite specifically on the Tyler model, taking into account the developments by writers such as James Popham and Elliott Eisner (1975). Eisner introduced a distinction between "instructional objectives: (where outcomes can be defined with some precision) and "expressive objectives" (where the task is more open-ended, as in aesthetic appreciation). Wiseman and Pidgeon asserted that teachers had unconsciously avoided the Tyler approach because of a preference for implicit rather than explicit goals. They go on to recommend an approach based firmly on the definition of aims in precise terms. The book edited by Tawney (1976) on the other hand, introduces a variety of additional approaches, arising from felt needs. These approaches echo the suggestions made by Hasting ten years earlier and establish evaluation procedures on historical, sociological and anthropological methods.

The source of this change can be quite clearly traced to the development of a large number of national curriculum projects, each of which required the appointment of an evaluator. Tawney lists thirty-seven such projects, featuring prominently the Schools Council and, to a lesser degree, the Nuffield Foundation and the Department of Education.

The Professional Judgement Model

Perhaps the most commonly used model of evaluation is that which might be termed the Professional Judgement Model of evaluation. This is the model that is used whenever evaluations are dependent on direct application of expertise or professional judgement, such as is the case in review groups empanelled by funding agencies to review proposals, site visits and evaluate federally supported programs, etc. It is difficult to trace the origins of this model, but the most clear cut and systematic example of this type of evaluation is school and university accreditation or certification of programs. It is also the most pervasive evaluation model in Australia and is commonly implemented through committee reviews or accreditation panels.

In accreditation, standards against which institutions or programs are measured are generally arrived at through collective judgements of persons seen as possessing appropriate expertise. The experts usually conduct a site visit and arrive at their final judgements after considerable observation and deliberation. However, criteria which are typically imposed on empirical data (eg. reliability and validity) are not imposed directly on information generated by this model.

Despite the common use of the Professional Judgement Model, it is so dependent on non-replicable mental processes of the experts who are asked to make the judgements that it almost becomes a pluralistic model.

The Ethnographic Model

A key paper in the British evaluation scene was the publication by Parlett and Hamilton, Evaluation as Illumination (1976). Their concern sprang initially from the comparative lack of success of many large curriculum projects both in the US and the UK and their feeling that the classical models of evaluation were inadequate.

Parlett and Hamilton in their approach show distinct similarities to the concepts put forward by Stake (1973), of evaluation as portrayal. The British authors draw a strong distinction between what they describe as the "agricultural paradigm" of evaluation, traditionally used, and their own preferred form, the "social anthropology paradigm". The approach depends heavily on participant observation and on the ethnographic field work of social anthropology. Other evaluators since these have pointed out the possible relevance of literary criticism, investigative journalism and film documentary. Parlett and Hamilton use the term "triangulation" to describe their three-stage approach: first, an overall study to identify significant features; second, the selection of a number of such features for more intensive inquiry; and third, the attempt at "explanation" through seeking general principles underlying the organisation of such a program.

Their work on illuminative evaluation was developed further by those with an interest in case studies, including Macdonald (1971), who was the evaluator in the Humanities Curriculum Project. In this project he made a quite deliberate and carefully considered move away from a psychometric approach to a style relying on methods usually associated with historical and anthropological research - for example, observation, interviewing and documentation. In addition, Macdonald redefined the idea of audience. The consumers of his work were decision makers of various kinds, the sponsors (the Schools Council and Nuffield Foundation), the employing authority, the schools, and the examination boards.

However, in the period since these initial publications on ethnographic evaluation first appeared, a number of doubts have been expressed about the so-called sociological or anthropological paradigm:

1. While there are established rules or procedures for anthropologists working in unfamiliar societies it does not necessarily follow that they can be carried over into curriculum evaluation.

2. The rules or procedures for non-traditional evaluation are insufficiently clear and the skills need to be specified more clearly. The area possesses no tradition comparable to the established standards of historical and anthropological research.

3. The variety of possible interests and audiences places evaluators in a situation of role conflict which may pose considerable strain on the evaluator.

4. There is a danger of subjective impressions being put forward as objective data.

5. The methods and language developing in this field may become as esoteric and remote from teachers and other interested parties as has conventional educational research.

These reservations do not dismiss this obviously important area but represent a series of problems which need to be solved. Worthen (1977) speaking from a wealth of practical experience as an evaluator as well as a wide knowledge of evaluation models, indicates that he had seldom, if ever, used a particular model as an entity in one of his own projects. "I couldn't think of a single lone-wolf evaluation of my own where I had consciously selected any single model to guide the study. Instead, for several years I have been designing each evaluation de novo, pulling pieces of the models in as they seemed relevant. Certain features of some models I used frequently, others seldom or never." He then goes on to argue that this is not so much a defect of current models as a reflection of a lack of a theoretical position with a sufficiently sound empirical base.

This viewpoint implies the need for a joint approach by those involved in evaluation, that is, combining the careful use of current evaluation procedures in a variety of contexts.

CHAPTER 3

LANGUAGE TESTING AND PROFICIENCY

Few areas in educational measurement have proven as complex or controversial as the testing of development in a second language. Approaches to the testing of a second language 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 preceding methods.

Current fashions in language testing have caused persistent difficulties for traditional measurement theory. Both the classical true score and error model, along with the correlational techniques of factor analysis, have been unable to deal adequately with many of the measurement problems in language.

This report describes the development of a series of objectives designed to generate a multitude of test items, a rationale for the organisation of the objectives and the application of a latent trait model to the problems of measuring second language proficiency. The objectives and the model are used as the basis for the development of an oral interview test. Further details of this test are supplied in the accompanying manual and technical papers.

Stages in Testing Methodology

Spolsky (1978) identified three major stages in language testing and this identification has been supported by subsequent authors (e.g. Farhady, 1979; Davies, 1982). The first stage identified by Spolsky, the pre-scientific stage, was typified by the grammar-translation approach in which little attention was paid to validity, reliability or the statistical properties of the test.

The second stage was identified as the psychometric-structuralist stage. This approach to testing, largely developed from the influences of Lado (1961) was typified by the use of discrete-point tests in which each element of the structure of language was tested using precise, objective tests. Duran describes them in these terms:

> Discrete point proficiency tests are composed of items. Each of these addresses an examinee's skill in controlling a single surface rule of the language related to morphology, phonology, grammar or vocabulary. (Duran, 1984:45)

The major aims in the development of tests of this style were measurement objectivity and reliability. Current thought, however, suggests that there may be serious doubts about the validity of such tests. The shortcomings of the discrete-point approach have been widely discussed elsewhere (e.g. Oller, 1979). The strongest opponents of this method point out that answering individual items, without regard to their function in communication, will not be of much value.

The third stage advocates the use of integrative tests which attempt to assess the skills involved in authentic communication. A rationale for these tests is expressed by Ingram (1984):

> "In real life all the components of language occur together, supporting each other in meaning and dependent on each other structurally; part of the skill of using language involves being able to put all the components together and to comprehend them when they are received together."
> (Ingram, 1984:3)

It is generally argued that integrative tests have greater content validity than discrete-point tests because they are more direct in their measurement of communication skills. While both cloze and dictation type tests have been generally classified as integrative because they do not isolate single component skills, they do not reflect "real life", and their format is regarded as artificial. This type of test has been classified as a pragmatic proficiency test, following terminology introduced by Oller (1979). The validity of cloze and dictation as integrative tests has been increasingly questioned over the past few years (e.g. Alderson, 1983). The most authentic and direct of the integrative approaches have generally required an interview that is rated according to its degree of accuracy, authenticity and acceptability.

The Notion of Proficiency

Along with the move towards the testing of authentic communication has been the development of notions of second language proficiency. A variety of definitions of proficiency can be found in the literature, the simplest of which regard proficiency as the learner's general ability to communicate efficiently in "real life" situations.

More elaborate definitions and explanations of proficiency have recently appeared in an American Council on the Teaching of Foreign Languages (ACTFL) publication, James (1985).

> (Proficiency is) a continuum, with isolated linguistic items at one end and individualistic language samples at the other, with a variety of combinations in between, their number limited only by the number of people using the language and the kinds of environments in which they operate. A person rated at Novice Mid in speaking, according to the ACTFL Provisional Proficiency Guidelines, has demonstrated a level of proficency that can be described and documented, whether or not a particular lexical item is present, or whether or not a particular structure is used in a certain way. (James, 1985:2)

> Proficiency is the outcome of language learning. It is not a method... Proficiency represents the basic principle upon which our profession has operated for centuries, namely, to help others control their personal and social environments by means of language and to obtain the greatest benefit from interaction with those environments ... such as the school and the street, the classroom and the boardroom, the casual conversation and the prepared speech, wherever it is possible to acquire or learn the skills of the language. (James, 1985:3)

At this point it is important to make a distinction between achievement and proficiency. The distinction is adopted in the discussion that follows.

Fundamentally tests designed to measure achievement and proficiency differ according to the kind of information they are intended to supply. Achievement tests determine the level of acquisition of specific course content. They are limited in scope and content and provide information about the extent to which a student has mastered particular material. Hence a student can study or prepare for an achievement test as the test reflects a specified body of material taught in a class.

Proficiency testing assesses a student's language performance in terms of the extent to which language is used effectively outside the body of material specifically taught in class. Proficiency testing should be curriculum free and is not concerned directly with where, how, or when a skill was learned or the exposure time taken for the student to develop the level of competence shown. Hence it should not be possible to prepare for the proficiency test. A proficiency test should sample language tasks independent of the specific instructional material. In this manner the proficiency test should find the limits of language beyond which the student is unable to progress at the time of testing.

Proficiency is an outcome independent of content, method, exposure time or location of acquisition languages. It represents the status of the student in terms of possible language usage at a specific point of time.

The Dimensionality of Proficiency

The dimensionality of language proficiency has been the subject of considerable debate over recent years. For example, a recent volume edited by John Oller Jr. is largely devoted to the examination of the proficiency dimension (Oller, 1983a). Papers in the volume approach the problem of proving a "unitary proficiency factor" hypothesis or a "divisible competence" hypothesis with the use of a number of factor analytic techniques. However, the very nature of factor analysis as a family of data-synthesising tools almost ensures that each author can find some evidence that supports his/her hypothesis. Further comment on the nature of many of these studies is made in Chapter 6.

An argument of this nature is unlikely to be solved through the application of factor analytic techniques alone. Despite this the authors argue strongly, on the basis of factor analytic results of varying quality, that their case has been supported by the data (e.g Farhady, 1983; Oller, 1983b; Vollmer and Sang, 1983). It is left to Carroll (1983) to restore some sanity to the argument:

> With respect to the issue of whether the data supports a "unitary language ability hypothesis" or a "divisible competence hypothesis", I have always assumed that the answer is somewhere in between. That is, I have assumed that there is a "general language ability" but at the same time, that language skills have some tendency to be developed to different degrees or at different rates, so that different language skills can be separately recognised and measured. A "general proficiency factor" is evidenced by the generally high correlations among a large variety of language competence variables, but one can also find evidence of specialised language skills in a more refined analysis of these correlations. (Carroll, 1983:82)

Arguments about "unitary" vs "divisible" skills will probably be fruitless and consequently a compromise of the type stated by Carroll is desirable. In order to begin an examination of dimensions in language proficiency, the first step is to define possible dimensions, construct valid and reliable measures of those dimensions and then examine the relationship between them. At the

individual test level what is required and must be investigated for the purpose of valid assessment is the validity of interpreting a total test score. At the individual test level a total score results from adding together scores on individual items. This score can only be validly regarded as an indicator of proficiency if the items "hang" together in some meaningful way. Similarly the student's scores on a battery of tests can only be pooled to form a single index of proficiency if they can be shown to "hang" together in a meaningful and interpretable way. As is evidenced by the literature, various methods of factor analysis have been unsuccessfully applied to the solution of these problems.

Validity of Proficiency Measures

The debate on the unitary vs divisible nature of proficiency cannot be resolved in this study. There are obviously many factors which impinge on the rate and nature of language acquisition. What is central to the controversy is the existence of a developmental dimension and the nature of that dimension. The argument seems to centre on whether a single measure can cover all aspects of development. Clearly it cannot. But whatever dimensions of proficiency can be identified, they must have two specific properties (Wright and Masters, 1982). The first is direction and the second is measurability. If language acquisition develops in a non-random fashion, then systematic development should be able to be identified and measured. The FSI, ACTFL and ASLPR scales support the argument that development is directional and that it is measurable as it passes through very broad stages.

The argument concerning other postulated dimensions should be examined with the same view. Factor analytic studies have assured us that direction can be identified, yet little evidence is put forward regarding measurement. For example, Johnston (1985b) postulates a second dimension, which he calls the variational dimension, but he offers little more than circumstantial evidence regarding a description of measures associated with the dimension.

In most cases the arguments for multi-dimensionality fall down on the measurability criteria. The evidence suggests that there are other factors to consider, but as yet only the proficiency or general development dimension is supported by evidence of directionality and measurability.

This is not to say that the other dimensions of language acquisition do not exist. It may be that until the problems of valid and reliable measures of the proficiency or developmental dimension are resolved, then attention and effort will be diverted away from a thorough definition of the other postulated dimensions and the controversy will not be resolved. The attention of this study is focused on improving the measurability of a dimension of proficiency. If the direction and measurement aspects of this dimension can be defined and refined, then the validity of the measure can become a point of departure from which other studies can address the issues of multiple dimensions.

In defining a dimension of proficiency, two considerations need to be borne in mind. First, we describe development in terms of probabilities. It is a normative development scale based on the expected development of a group. Specific individuals can be expected to differ as Johnston (1985b) illustrates with his variational dimension. It is groups of individuals who develop in a systematic fashion. The instrument used to define the dimension then must have sufficient structure to avoid confounding influences. Second, where the deviations from the expected direction and rate of development are systematic across groups, other influences such as first language, psychological, social, and educational factors will have to be explored to determine whether these can aid in explaining the departures from the expected systematic development.

While many studies have examined the existence of proficiency factors in test batteries, examinations of the proficiency dimension as described in the ILR, ACTFL and ASLPR scales are hard to find. What is needed is a thorough empirical examination of the proficiency dimension as described in these instruments through the application of a sound theoretical model. Bachman and Palmer (1983) used a multi-trait, multi-method approach along with confirmatory factor analysis to examine the construct validity of the FSI scale. They found a strong method effect for the FSI, but it was shown to be less than the method effects of a set of self-ratings and translation exercises. They also argued that their data supports a divisible factor hypothesis. In the results of the formal trialling of the ASLPR, Ingram claims that:

> "Construct and content validity should have been assured by the development process in which the scale descriptions drew on psycholinguistic research and were assessed in numerous face-to-face interviews." (Ingram, 1984:19)

This is hardly sufficient to demonstrate construct validity. In trialling, Ingram demonstrated that the interview results based on the ASLPR have some concurrent validity with other ESL tests and claims that its widespread acceptance is an indicator of its face validity. It is, however, much more difficult to demonstrate construct validity and classical testing approaches are not well suited to this type of validation. Furthermore, the claim that a dimension of proficiency exists and is adequately described by the ASLPR means that construct validation is essential.

CHAPTER 4

A MEASUREMENT MODEL

The Need for a Measurement Model

From the arguments specified above it is possible to identify the requirements of a measurement model that can improve the development of second language tests and their interpretations. Current thought suggests that oral interviews are the most valid means of measuring second language oral proficiency. In programs developed for the selection, screening, or placement of students according to oral proficiency language it is typical to have large numbers of candidates involved. Since the criteria of proficiency is so complex it is difficult to construct valid, reliable, direct measures which can be routinely scored. The complexity of the task has generally resulted in the use of an impressionistic, unstructured interview that may be followed up with discrete-point tests of component skills.

The lack of reliability in the interview and the potential lack of validity in the discrete-point test negate the effectiveness of the combined exercise. The strength of one exercise does not compensate for the weakness of the other.

Validity is enhanced by directness of measure. Tests of proficiency therefore need to involve authentic communication tasks and would seem to require the interview format. Reliability is enhanced (but not assured) by routine scoring procedures. Hence the interview needs to be structured so that it can be scored objectively. The use of conversational or interview techniques also require one-to-one approaches. The test therefore would need to be short to increase efficiency while retaining validity and reliability (Griffin, 1985).

Classical test theory has not proved to be adequate for the design and calibration of this type of test. In particular there are problems with the classical notion of scoring test items dichotomously (right/wrong). In an interview it is far more natural to grade a student's response in a number of categories depending on its degree of acceptability. Secondly, the data provided from a holistically scored oral interview does not lend itself to easy use for the diagnosis of the particular problems of individual students.

In the following sections a measurement model will be outlined. This model can be used to explore the existence of a proficiency dimension and the meaningfulness of interpreting a global score of proficiency. If such a dimension exists it can then be used to provide reliable estimates of a student's ability when used with interview based tests and items that require the use of polychotomous scoring. It can also be used to diagnose specific areas of weakness in a student's language proficiency.

Defining a System of Measurement

We can use an analogy put forward by Choppin (1982) to establish the setting for rethinking the approach to the measurement of language proficiency.

The way in which such physical measurements as temperature are obtained, can serve as a model for the measurement of human abilities including proficiency in spoken language. Consider the fact that the temperature of a substance is independent of the means of the measuring device used. If the device is sufficiently accurate a consistent reading should be obtained. Furthermore, all substances at the same temperature should register the same reading. These two aspects are worth considering. The temperature is independent of the measuring instrument and the measuring instrument is independent of the substance being measured. The reading obtained is a result of the temperature of the substance and the gradations on the measuring instrument. It could be, and sometimes is, argued vociferously that measuring human abilities is a different concept from the measurement of physical entities. This is clearly true for abilities which do not lend themselves to a direct approach to measurement. We cannot use a ruler, thermometer or stopwatch to measure a person's language proficiency. We can, however, define a range of tasks from the domain of language skills and observe his/her performance. From these observations we can make inferences about proficiency. The measurement is indirect. The ability or proficiency has to be treated as hidden or latent.

This should not prevent us from seeking an approach to measurement which contains the same desirable elements as the system used to measure physical entities. That is, the ability being measured should be independent of the measuring instrument, and the actual measurement should not depend on the person(s) being measured. Measurement instruments should be interchangeable and it should not matter which proficiency "thermometer" we use. A measurement system is required with specific properties.

1. The measure obtained on the instrument should be independent of which substance (person, sample, population) is being measured. The measure obtained should only depend on the amount of trait or ability present.

2. The instruments should be able to measure a large range of abilities. That is, the "gradations" should cover more than a narrow range but be sufficiently close together that some fine-tuned measures are possible. The instrument should not be affected by factors other than the trait it is designed to measure.

3. The range of instruments designed to measure the trait should be interchangeable. It should be a matter of indifference which instrument is used to obtain the measure. Instruments capable of cross-calibration are required.

As a consequence of these three properties, the knowledge of a person's score on one test would make it possible to predict the score on another test. Traditionally this has been achieved via norm-referenced techniques using percentile ranks or standardised scores within the same sample or population. The reference population must be defined in order to interpret the score. Hence the measure obtained and the system of measurement fail on the first of the three measurement criteria. The use of percentile ranks means that the gradations between the score levels are not consistent across different samples or populations. Additional gradations such as age-norms, cultural indices or national norms need to be introduced as a separate percentile or gradation scale developed for each group even with the same instrument. Hence the instrument and the measurement system fail to meet the second requirement of a measurement system. Clearly, interchangeability of instruments is only possible within defined populations and is not possible outside those defined groups. The norm-referenced system thus fails to meet the third criterion and does not provide a solution. A measurement system is required which approximates the techniques of physical measurements.

The Rasch Family of Measurement Models

During the 1950s and 1960s a Danish mathematician, Georg Rasch, began the development of a family of measurement models that satisfies the criteria specified for a measurement system as described above. The use of these models has been described in a recent paper by Griffin (1985).

The most basic of these models is the simple Rasch dichotomous model (Rasch, 1980; Wright and Stone, 1980) for use with test items scored right/wrong. In this case every person is described by one parameter (ability) and every item is described by one parameter (difficulty). Whenever a person attempts an item, the ability and difficulty parameters interact to give the probability of a student's response to the item. The mathematical form of the model is given by:

$$\Pi_{ni1} = \frac{exp(\beta_n - \delta_i)}{1 + exp(\beta_n - \delta_i)} \qquad (1)$$

Where Π_{ni1} is the probability of a person with ability β_n scoring one on an item with difficulty δ_i.

Figure 1 shows a graphic presentation of this interaction between item difficulty and person ability. These curves are called item characteristic curves.

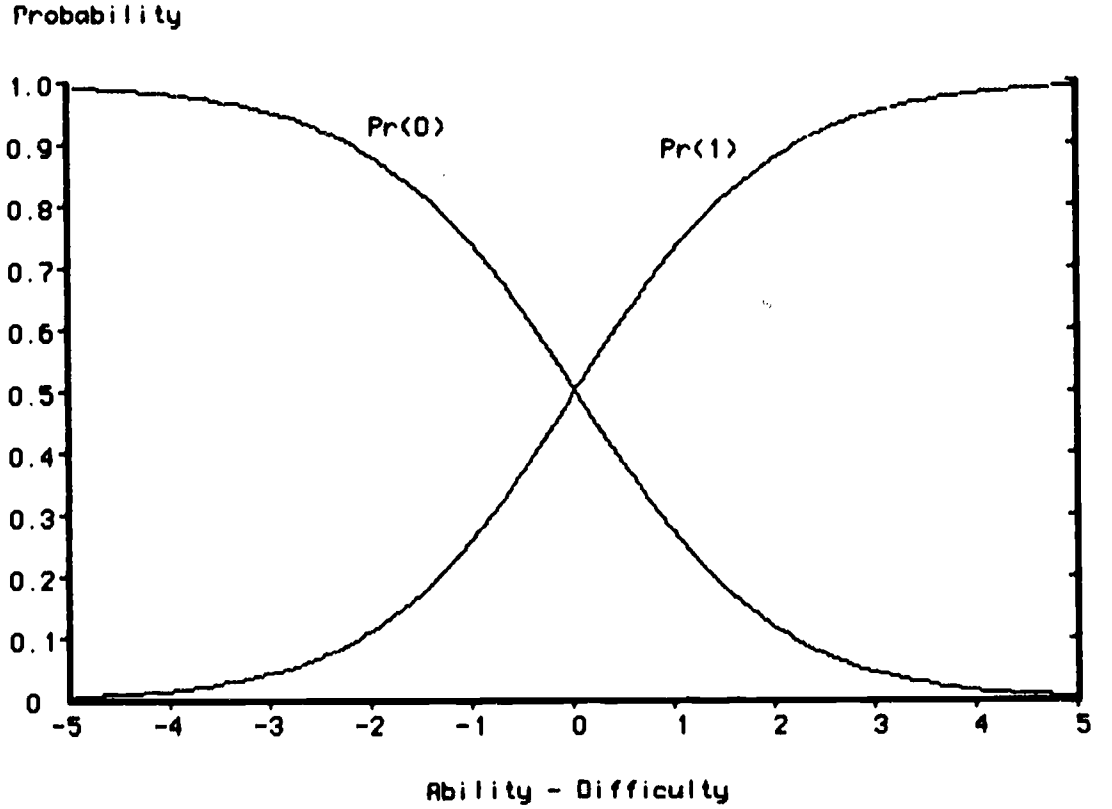


Figure 1: Item Characteristic Curves for a Dichotomous Item

The probability of the score one (correct) or the score zero (incorrect) on an item can only vary between 0 and 1 and this is shown on the vertical axis. The horizontal axis shows the difference between ability and difficulty. As the difference between ability and difficulty increases, the probability of a correct response decreases to zero. In the Figure, the difference between β_n and δ_i is zero at the point of intersection. At this stage the probability of a correct answer is 0.5. That is, there is an even chance of a correct or incorrect response. To the right of this point, the person ability is greater than the item difficulty and the probability of a correct answer is greater than that of an incorrect answer. Similarly as difficulty exceeds ability the probability of a correct response decreases to zero. In this region (to the left of the intersection point) the probability of an incorrect response is greater than the probability of a correct response. The form of the model makes intuitive sense since as ability increases the probability of success increases.

Statistical properties of this model allow for the separability of the item and person parameters. This means that it is possible to make the person ability independent of the measuring instrument (or the difficulty of the items) and item difficulty becomes independent of the person or group to which the test is administered. This single property ensures that the model satisifies all three criteria of the required measurement system as outlined above.

Another member of the Rasch family of measurement models is the Rasch Partial Credit Model (Masters, 1982). This model has all of the properties of the required measurement system but it also allows for the scoring of items in more than two categories. This is a necessary requirement of the structured interview format. Essentially this model is an extension of the basic model described above so that items can be scored in any number of ordered categories. For example, a student's response to an item may be graded 0, 1, 2, or 3 according to its degree of increasing acceptability. That is, there are varying degrees of correctness rather than the totally correct or totally incorrect classification allowable with the dichotomous model.

Under this model the probability of person, n, responding in category x to item i is given by:

$$\Pi_{nix} = \frac{\exp \sum_{j=0}^{x_i} (\beta_n - \delta_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} (\beta_n - \delta_{ij})} \qquad (2)$$

where x takes the values 0, 1, 2, ...m_i corresponding to the $m_i + 1$ ordered categories associated with item i, β_n is the ability of person n, and $\delta_{i1}, \delta_{i2}, ..\delta_{im_i}$ are the m_i difficulty parameters associated with item i.

For notational convenience $\sum_{j=0}^{0} (\beta_n - \delta_{ij}) \equiv 0$

Item characteristic curves for items with varying levels of correctness, using a score procedure of 0, 1, 2, 3 instead of 0, 1, can be drawn on a similar basis to those shown in Figure 1. Figure 2 shows an example where four separate categories of scoring are used, i.e. 0, 1, 2 and 3. As in Figure 1 the probability of a particular score is shown on the vertical axis and the difference between ability and difficulty is shown on the horizontal axis. To the right on the horizontal axis it can be seen that ability exceeds difficulty and the probability of responding in the higher categories increases and the probability of scoring in the lower categories decreases. At all points along the horizontal axis the sum of the curves must add to one since the probability of scoring one of the four possible scores must add to one.

The model is used to estimate the probability that a person will make a particular response to a test item. Unlike norm-referenced approaches with traditional tests it is possible to express the ability and difficulty parameters in the same units, thus enabling the direct comparison of the person's ability level with the difficulty of the task. As ability increases the probability of a higher score increases, and as ability decreases the probability of an incorrect response increases.

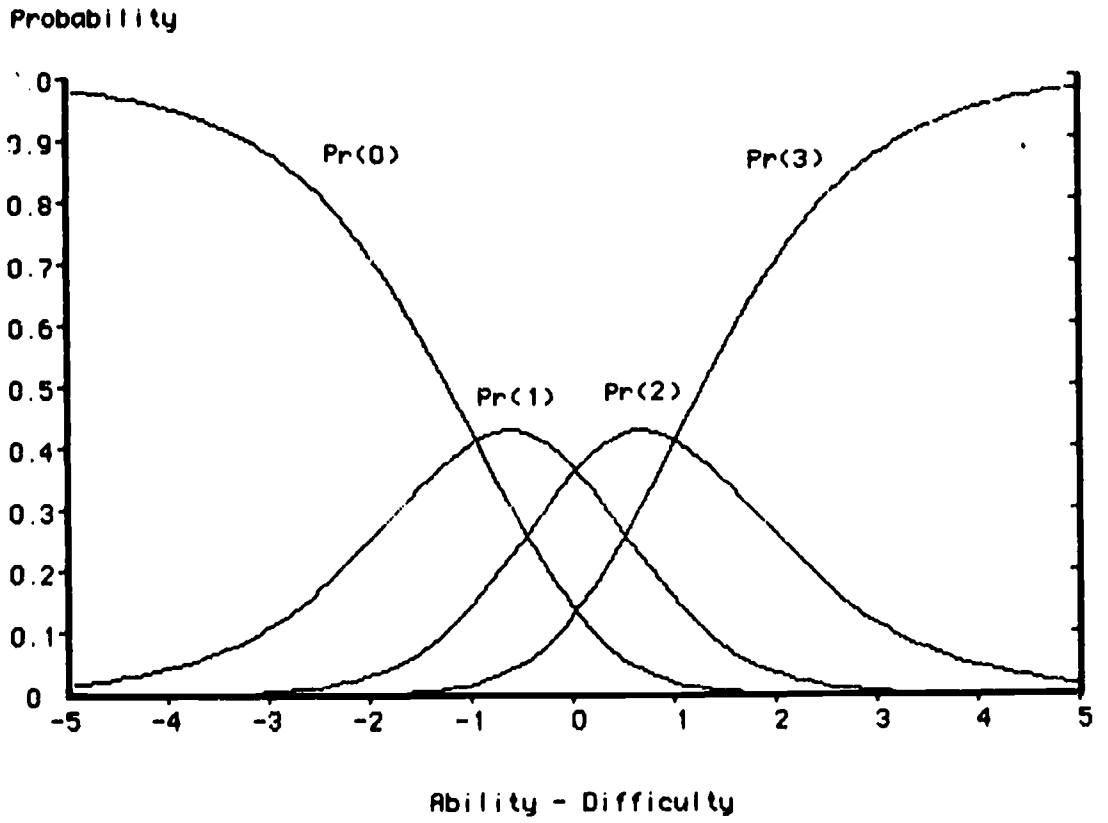


Figure 2: Item Characteristic Curves for a Polychotomous Item

Assumptions of the Model

The specifications of the form of the Rasch model in (1) and (2) imply strong assumptions that need to be carefully examined wherever the model is applied. First, since each person is described by a single ability parameter and each item step is described by a single difficulty parameter, a unidimensional latent trait is assumed. Consequently in the case of second language proficiency, it is assumed that proficiency develops along an orderly continuum. Clearly with such a complex skill as language this may not be the only continuum, but one of many as yet undefined dimensions of development. Due to the complexity of language proficiency we may find that the model is influenced by other dimensions but, if this were the case, we would have also rejected the existence of a measurable proficiency dimension that is so widely supported in the second language teaching/testing community. Secondly, the use of a single parameter for each item step implies that each item is independent of all other items in the test. This assumption of "local independence" means that a student's performance on one item should not be affected by his/her performance on other items in the test. This assumption

may be of particular concern in an interview format if the interviewer corrects a student when he/she makes an error, thereby altering the student's expected performance on subsequent items. The assumptions of the model are generally examined through indices of fit. These indices measure the degree to which the data conform to the model specifications. The data to model fit should be carefully examined during any test construction and calibration exercise to ensure the validity of the approach,

Properties of the Model

There are critics of the model. Arguments against the application of the Rasch model are based fundamentally on the rigidity of the assumptions, the arbitrariness of the scale and the tests of fit (Goldstein, 1979). The model does make strong assumptions. However, it has been shown to be quite robust to violations of these assumptions (Izard, 1981; Andrich, 1982), and in situations where the assumptions are clearly not met, a measurement specialist would not apply the model. Objections to the arbitrary scale seem a little strong given that all measurements of latent human abilities are arbitrary regardless of the measurement model. A score of 23 out of 30 on any test is a purely arbitrary one, as are percentage measures or even the classifications of the FSI, ASLPR or ACTFL materials.

As Goldstein (1979) correctly pointed out in his criticism of measures derived from the application of the Rasch Model, they form an arbitrary scale. But it does possess proper measurement qualities -- something that conventional (arbitrary) scoring procedures cannot provide -- and the arbitrariness of the measure should be no more of a handicap than that of degrees used to measure temperature. The raw scores of gradations may differ from one test to another, as may the position of an arbitrary zero point. In applications of the Rasch model, the zero point is centered about the mean as a convenience only. However, each test constructed to measure the same trait can be cross-calibrated and ability measures can be equated just as centigrade and Fahrenheit temperatures can be equated on to one common scale.

Individual interview items can also be mapped on to the common scale and it is this property of the model which enables libraries or banks of test items to be developed.

From such stores of items, short tests can be quickly developed to measure a trait, with accuracy, near a suspected level of ability. In any one centre, this item banking could provide language measurement specialists with pools of test items to enable precise measurements to be made which would be directly comparable to ability estimates anywhere else. Detailed discussions of the calibration, equating and item banking procedures go beyond the scope of this paper but excellent descriptions may be found in papers by Wright and Stone (1980), Wright and Masters (1982), and Masters (1984).

The test of fit of data to the model is being given considerable attention. The basic difficulty seems to be that there is no single sufficient test of fit. It is necessary to apply a series of tests. The objections to the use of the Rasch model were based on the application of a single test and such objections are well founded but should not be directed at competent users. An excellent discussion of the tests of fit can be found in a paper by Douglas (1982).

The measurement model does have limitations. It would be folly to pretend otherwise. However, within its limitations, it still makes a great deal of sense to utilise the separability of parameters it offers. The benefits of this property far outweigh any costs involved and greatly exceed the potential of conventional test construction and calibration procedures.

Despite the complexity of the formulae presented earlier, the model is mathematically quite simple and convenient. Estimations of the separate parameters do not require large amounts of data. The only necessary and sufficient information for estimating the parameters is the total score over all items for each person, and the total score over all persons for each item. It is these conditions which characterise the Rasch model. A number of estimation procedures have been developed and discussion of them can be found in Wright and Masters (1982). It is sometimes called the one-parameter model because in its simplest form only one item parameter is estimated. However this terminology leads to misunderstanding. It is possible to obtain more than one item-parameter estimate using the Rasch model (Andrich, 1978a, 1978b, 1978c, 1982; Douglas, 1978; Masters, 1980, 1982; Wright and Masters, 1982). The distinctive property of the Rasch model is the separability of the person and item parameters and the limitation of the necessary and sufficient information to the description of person and item scores. No other latent trait models possess these characteristics. No normative model provides the independence and specific objectivity of scores.

In the past 20 years sufficient work has been completed on the model to demonstrate that it can and does predict the behaviour of real people and real test items with consistency and accuracy. Other, more complex, models (Birnbaum, 1968; Samejima, 1969; Bock, 1972; Lumsden, 1976) present persistent difficulties in obtaining estimates. Since the only information we can use is the person and item scores, models which do not allow the complete separation of the parameters have been shown to be overparameterised (Andrich, 1982). This means that it is not possible to obtain sufficient data in a testing situation to fully fit the data to the models. Often the researcher must make estimates of guessing, discrimination etc., in order to use more complex models. It has also been assumed that the Rasch model requires constant discrimination of items, but this has also been shown to be false. The equality of discrimination of items is a result of fitting items to the model in the dichotomous case and is not a general requirement of the model (Andrich, 1982). The only necessary aspect of the Rasch model is that the parameters are separable and are based on the descriptive data of persons and items.

CHAPTER 5

ORGANISING PRINCIPLE

Structured interview tests are not new, nor is the application of latent trait calibration to such tests. In fact Rasch originally developed his model to assess the reading ability of Danish school children using a verbal test.

In the following example we attempt to show how the model can be used to solve many of the measurement problems associated with testing oral language proficiency. The application of this model allows us to develop a scoring scheme that contains a considerable degree of objectivity yet does not rely on scoring in only two categories (right/wrong). Objectivity will allow for greater reliability and the opportunity to give partial credit through scoring in a number of response categories will allow for more scope and indeed we believe more validity in the rating of authentic language performance. The benefits in terms of diagnosis of specific language difficulties are also evident from our data.

The application of this model allows us to investigate the existence of a dimension of language proficiency. There is some agreement that such a dimension exists and that second language learners develop along this proficiency dimension from "zero proficiency" to "native like proficiency". If the data fit the model, then we cannot reject the notion of a dimension of proficiency. Under these circumstances, we can use the data to examine in detail the performance of individual items and persons on the assumed underlying dimension. It can be also argued that a single index of proficiency has meaning, provided that the items used to define the dimension have interpretable, ordered sequence. It is this evidence that is used to define the nature of the dimension and to argue for its validity.

Given the possibility of a points earned or a partial credit approach, the generalisation of the Rasch model can be applied to sets of language tasks, each scored on a routine rating scale. The simplest of these could contain the ratings categories "wrong", "partly correct" and "totally correct".

The first step, of course, is to develop tests of proficiency. This type of test should reflect authentic reactions to language situations experienced by the student. Proficiency should not be assessed in isolation from the natural use of language. The assessment should serve the needs of both students and teachers as they plan, shape and reshape the learning and experiences to satisfy needs. Because of this, the definition of proficiency and the

application of assessment will shape the curriculum, and vice versa. The difficulty is that the strict isolation of the testing situation is hard to avoid. A partial solution is to develop the assessment around a natural-like discourse and dialogue.

Standard test development techniques suggest two promising approaches:

1. The curriculum analysis approach.
2. The actuarial approach.

The curriculum analysis approach assumes that a detailed analysis of the spoken language components has been conducted. These components are translated into test specifications and the items developed from these. A full description of this approach to test construction can be found in the outline by Burrill (1976) of procedures for developing standardised tests.

There may be some argument regarding the suitability of the curriculum analysis approach since proficiency should be seen as curriculum free in nature. However, an analysis of curriculum should identify a series of general skills necessary in the development of proficiency. The tasks or techniques used to assess the development of these skills should be curriculum free. That is, the students should not have had the opportunity to practise the material. Under these circumstances the test becomes more an achievement test than a general proficiency test. Nevertheless, a series of ordered objectives is still required in order to define the general skills and to define the nature of the assumed proficiency dimension.

The actuarial approach also has something to offer developers of language tests. The actuarial approach to test construction, outlined by Cronbach (1970) demands that persons highly skilled in the desired attribute (e.g. spoken language) are identified and their observable characteristics noted (relevant, of course, to the trait in question). That is, how would a native speaker respond to a given situation or use language effectively? The same procedure can be adopted across agreed levels of language competence such as those outlined by the ASLPR. An example of this technique can be found in Johnston (1985a) and is discussed more fully in later sections.

This approach means that not all student performances are compared to a standard of native proficiency. While there may be some disagreement about what constitutes native proficiency, this ideal provides a more soundly based approach to measures of proficiency, in that the authenticity of performances may be used as a possible criterion for defining mastery of each developmental stage.

Elements of both the curriculum analysis and actuarial approaches were used in this study. In an examination of current literature, an examination of existing materials, observations of classroom lessons, interviews with teachers and in-service activities, a curriculum analysis approach was used to identify elements of language proficiency that were considered most important. These elements were used as the basis for the development of a set of objectives. The actuarial approach was then used to develop performance criteria for the objectives and example test items.

CURRICULUM ANALYSIS

In the identification of an ordered series objectives for the purpose of assessment of proficiency in a second language, a search for what is "necessary and possible" was undertaken. This search began at the theoretical level with an investigation of recent developments in linguistic theory.

Developments in Theory

Since the publication by Wilkins in 1972 of what has become to be known as the Functional/Notional Syllabus, the language teaching profession has been in turmoil. He suggested that the structural principle was insufficient for instructional design, as it was inadequate in terms of defining the uses of language. He therefore proposed two categories of "meaning" and "use" which might be suitable for the purposes of instructional design. The first category, which he calls "semantico-grammatical", expresses such ideas as "frequency, duration, quantity, etc." These categories are now classified as "notions". Whilst these notions are items of meaning, they relate fairly directly to grammatical categories in European languages. Wilkins's second category is the "communicative function". Communicative functions are the broad uses to which we put language. "Functions" express such uses as "requesting information", "giving orders" and so on.

If the structural ordering of the language derived from the classical tradition of teaching hierarchically from the simplest items to the most difficult needed to be re-assessed in terms of adult practical needs for communication, grammatical ordering in the old way had to change and some sort of ordering based on meaning had to take its place. It was thought that it would be possible to organise a syllabus based on Wilkins's ideas.

However, as course developers found, the resultant definition of objectives in these terms tended to develop arbitrary lists of functions and notions and contradicted a characteristic inherent in language use, i.e. the capacity to react appropriately to things which cannot be for seen or defined. It was also not clear that a knowledge of language expressed in functional/notional terms enabled the user to generate new sentences unaided. Wilkins wrote:

> The grammar is the means through which linguistic creativity is ultimately achieved and an inadequate knowledge of grammar would lead to a serious limitation on the capacity for communication. (Wilkins, in Page, 1979:117)

> It is taken here to be almost axiomatic that the acquisition of the grammatical system of a language remains a most important element in the language. The notion that an individual can develop anything other than a rudimentary communicative ability without an extensive mastery of the grammatical system is absurd. (Wilkins, 1981:93)

Whilst the communicative dimension in course development writing is now accepted as an essential component, the problem has centred on how to combine these ideas into a workable classroom model. It would seem that Brumfit (1981) has found an acceptable solution and that solution may yield one method of analysing syllabi for a common generic and ordered series of language tasks.

> The syllabus will be specified grammatically, because syntax is the only generative system so far described for language, and - since time is at a premium - a generative system will be more economical as a way of organising language work for student learning than a non-generative taxonomy of items (such a list of functions is at the moment bound to be), or a random selection of items, unsystematically collected. However, any attempt to contextualise or situationalise the grammatical items will

> involve a variety of language functions being used and a variety of notions being realised. It will not be difficult to bargain appropriate functions or notions (if they appear to be unduly neglected or omitted altogether) against the syntactic forms being used. That is, the ordering of items in the syllabus will be determined by a cross-fertilisation between functional and grammatical categories, but with the generative system fundamental. You could thus conceive of the syllabus as a grammatical ladder with a functional-notional spiral around it. (Brumfit, 1981:50)

Whilst Brumfit goes on further to add that such syllabi do not overcome Wilkins's objections to a static view of language learning as might be entailed in a structural syllabus, the development of fluency or communicative competence in the classroom may help to overcome this problem.

Developments based on proficiency must allow students to make grammatical mistakes in the development of communicative competence during instruction. These must also entail an input in which correct structures are taught. As real communication in the spoken language takes place in real time, structural errors are inevitable before the mastery stage is reached, but the communication stage may be still present. Proficiency testing, however, is concerned with the assessment of mastery and identifying those areas where communication occurs despite the pressure of structural errors. It is in these areas that perhaps the greatest and fastest growth in development may take place, leading to proficiency gains. Accordingly it seems that linguistic accuracy should not be postponed as such an approach promises a terminal profile.

> It is scarcely fair to lead serious students along encouraging them to talk like Tarzan and then saying that future progress is unlikely if not impossible until they get their grammatical act together. (Higgs, 1984b:7)

Whilst a grammatical approach as presented in the classical tradition is currently looked upon as stultifying, recent writers have pointed out not only the generative aspects of structure but also its meaning potential. Students need to learn not a number of finite speech acts, but rather structures which will suffice for a number of functions: Widdowson's instruction, invitation, advice and prayer are all expressed in the imperative form, illustrating the diverse applications, communication and intent of a simple grammatical structure:

"Bake the pie in a slow oven.
Come for dinner tomorrow.
Take up his offer.
Forgive us our trespasses."
(Widdowson, in Paulston, 1981)

Higgs makes the same point in these terms:

> It is not popular these days to stress the need for grammatical accuracy in foreign language classes. But it is vital to recognise that once a fairly elementary level is surpassed, the grammar itself communicates considerable meaning. It makes a great deal of difference in English to say - if you and I are friends we can discuss this openly - when you mean - if you and I were friends we could discuss this openly. (Higgs, 1984b:7)

One major area of contention still remains unresolved. Traditional grammatical models, taught as a series of rules and procedures have not only been criticised as stereotyping but recent work even questions the validity of the ordering of these particular grammatical models. While it is possible to leave the proficiency debate as unresolved, it is important that an appropriate grammatical model be agreed upon if the assessment of proficiency is to be based on the syntactic generative system. This is to say the validity of proficiency measures is questionable if an appropriate grammatical model is not employed. Without such a model it is possible to teach for communicative competence allowing for, and perhaps expecting, grammatical errors, but if proficiency assessment is to have any meaning an appropriate grammatical model needs to be identified.

The cornerstone of the proficiency movement as developed in the USA concerns the trisectional model of function, content and accuracy. Based on developments of the FSI and experience of proficiency testing on the ILR Scale, the American Council for the Teaching of Foreign Languages (ACTFL) has developed generic descriptions for language skills based on a scale of developing proficiency. In each new level in the ACTFL/ILR definitions, function, context and accuracy are grouped together in a constellation of relationships. It is argued that it is inappropriate to specify any one element of language development as the over-riding influence. In demonstrating this interrelationship among the elements of language development, Higgs and Clifford (1982) have developed a relative contribution model depicted in Figure 3 below. The figure illustrates most of the accepted

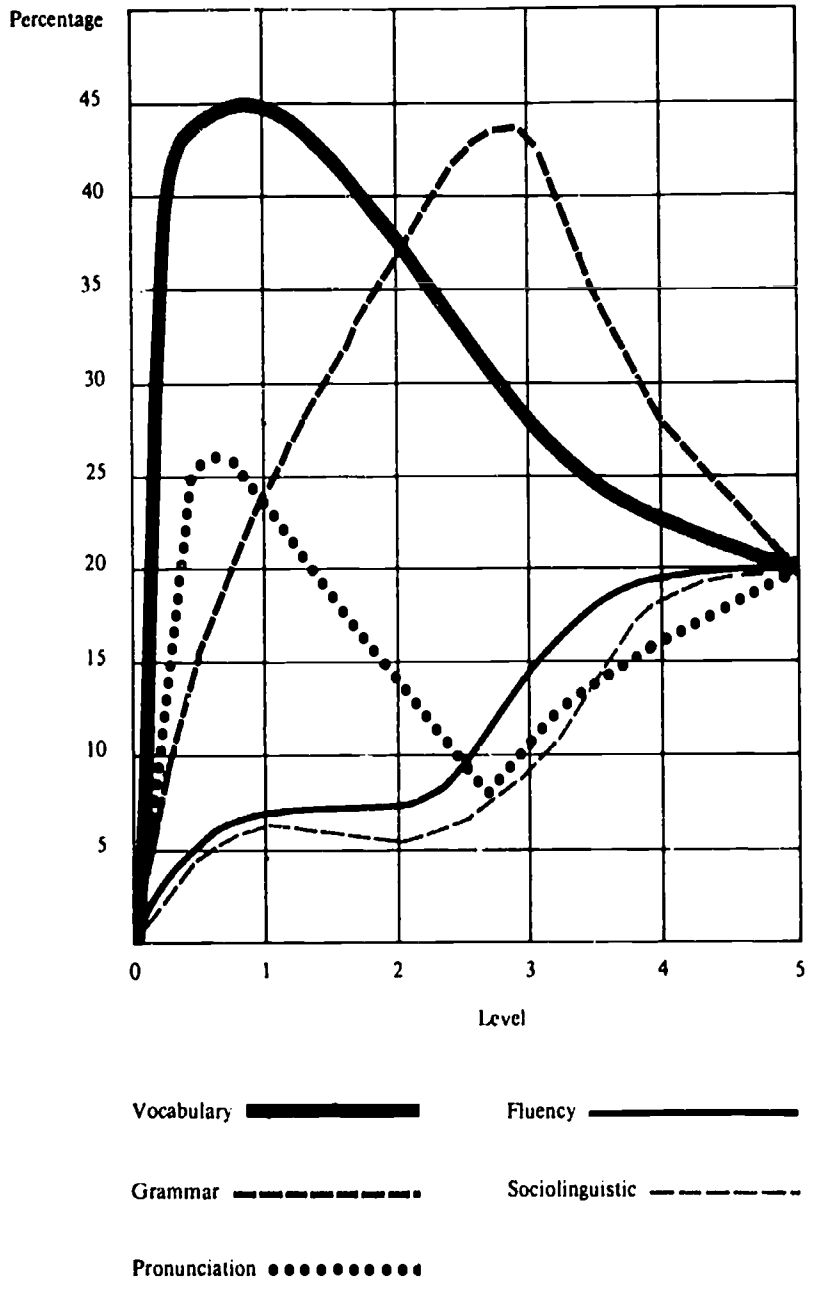


Figure 3: Hypothetical Relative Contribution Model

notions of proficiency and indicates that any generic model should follow a particular sequence of development. The relative contribution of specific language subskills - pronunciation, vocabulary, grammar, fluency and sociolinguistic appropriateness - changes from level to level.

It is essential to remember that the height of the curve at any given proficiency level indicates the contribution for each subskill. The fact that the vocabulary curve drops as it approaches level 5 does not mean that less vocabulary is needed, but that in comparison to the other four contributory

skills vocabulary declines in relative importance. Because this is a graph of relative contributions the values of the five curves at any given level always total 100 per cent. At level 1 the most important contribution is proposed to be made by vocabulary, followed by sufficient grammar to create with the language and a minimum threshold level of pronunciation sufficiently accurate to be understood. Fluency and sociolinguistic elements are not yet crucial, because at this level one is concerned with listeners who are used to dealing with foreigners, and the expectations of both the speaker and the listener are quite low.

According to the proposed model, at level 2, these relationships would shift. The relative contribution of grammar would increase, as the required linguistic task (i.e., the range of linguistic functions to be mastered) became more complicated. At the same time, the relative importance of pronunciation would begin to decline after reaching the minimal level required to be understood.

The relative contributions of the subskills as graphed, peak at different levels. For example, sociocultural variables, which play a minimal role at the low end of the rating scale, are crucial to attaining an ILR rating of 5.

As Higgs points out, the Relative Contribution Model has serious implications for any discussion of instructional methodology and, it may be assumed, for the definition of an ordered set of generic objectives used in defining a general proficiency dimension. The novice level (i.e., ILR-0), for example, implicitly recognises "enumeration of memorised material" as the primary function that can be expressed. Content is concerned "with common, isolable, semantic groups of lexical items such as basic objects, weekdays, months, meals etc. Accuracy is limited to intelligibility. Not surprisingly, at the novice level, the vocabulary and pronunciation curves of the contribution model are near their peak. If this level is the first step up the proficiency ladder, debate over the merits and liabilities of the grammatical versus the functional/notional syllabus are moot. What is needed, is to focus on a lexical base. Nothing else could reasonably be asked of the students in order that they may progress:

> Since the grammar curve nears its peak at the 2+ level, proficiency theory predicts that a grammatical syllabus is not only defensible, it is indispensable. (Higgs, 1984b)

This, then would need to be the main but not only focus of proficiency at that level. Which aspects become the focus of proficiency assessment, however, depends on two things. First, the tasks have to be identified, defined and ordered on the proficiency dimension. Secondly, a means of identifying the acceptability of accuracy of the performance by the student has to be defined. That is, performance criteria have to be defined. The ordering of tasks, and the definition of performance criteria are interrelated. But the first step was to validate the proposed contributory model.

In order to verify or refute the model Higgs and Clifford asked fifty language specialists representing seventeen of the languages taught in the CIA Language School to identify and rate the relative importance of the contributory skills for each proficiency level on an instrument that paired each subskill with every other subskill. All of these specialists were members of the Language School staff and were familiar with the proficiency level definitions. Rankings made by teachers supported this hypothesis and scale values developed from the paired comparisons in a range of languages came even closer to the hypothesised model (Higgs and Clifford, 1982). The same exercise, conducted with teachers employed in the AMES in Melbourne, yielded very similar results, indicating that the Higgs and Clifford profile was also applicable to AMES classes and hence to their assessment.

An Examination of Materials

In attempting to define objectives for language proficiency on which testing material for this project could be based, a review was made of a number of commercially available testing materials along with resources available to the AMES teachers.

A search of materials available to teachers reflected changing perceptions in second language acquisition. Early documents, including class work sheets, reflected a discrete point approach to the testing of structures. These had a written orientation.

A search through the resources available to teachers revealed a wealth of material although much of it of mixed quality. The relevance of these materials also reflected changing perceptions in second language acquisition. The number of these materials makes it impossible to comment here other than a few of the most widely used or most relevant.

The Michigan Tests of English Language Proficiency were considered to be too difficult for the testing of the students with whom this project was concerned, as the level of language required, for example, in aural comprehension was beyond the comprehension of students with very limited English. The Royal Society of Arts Tests were recently developed to test oral language. The tests are task based, but were again considered far too demanding for students whose production might be limited to monosyllabic, verbless utterances. The Ilyin tests (1976) were seemingly more suitable for our purposes, but they too demanded a proficiency above the level of many of the AMES clients. Furthermore, it was felt that the testing style of some aspects of the interview was unacceptable to teachers as an authentic testing format.

In the examination of commercial materials a testing schedule developed by the ACTFL and ETS was found to be of highest quality and relevance from a theoretical perspective. The theoretical underpinnings of this interview were based on the FSI and the work of Higgs and Clifford (1982). Their approach had been developed on the basis of the relative contribution model shown in Figure 3, but it was focused at higher proficiency levels than those most common at the AMES. The theoretical underpinnings of this material gave considerable support to the research team in the approach that they had begun to adopt as a result of examining the literature.

The interview-based testing materials identified at the AMES were largely designed to supplement the ASPLR. Of these the most highly developed was a package called the ASLPR Kit developed by teachers within the AMES. It provided a framework for organising and directing an interview that could be used to globally assess a student's ASLPR. Some of the material in this kit was drawn upon in the development of both the pilot and final testing materials.

An extensive project undertaken by the Australian Council for Educational Research (1974-1976) under a commission from the Australian Department of Education resulted in the Tests of English for Migrant Students (TEMS Tests), a comprehensive set of eighty-four tests and checklists designed to measure the competence of migrant students in various aspects of the English language (ACER, 1976). The authors describe the tests as follows:

> Each test has been designed to give information about a single aspect of the student's performance. Such information can be interpreted as being information about the student's underlying competence in the language; about, that is, ability to understand and produce symbols or sounds, the words, the syntax and the meanings which constitute the English language as used in Australia. The student's performance on a test is to be interpreted in terms of an applied criterion for that test. Each criterion can be expressed in terms of a particular skill, and implicit in each test is the belief that every competent user of the English language in Australia has the ability and should have the facility to use that particular skill. The level of facility, or the degree of accuracy with which a skill is used, varies among migrants (as, of course, it does among native speakers of English); it also varies according to ethnic group, length of stay in Australia and frequency of exposure to English. (ACER, 1976:3)

As the developers of these tests point out, at the time of their construction, there was no clear model of second language acquisition which could be used by the test constructors. Recent work in this area by Pienemann and Johnston (1984) would indicate that this problem is now being overcome and this research will be discussed later.

Whilst the contexts of the TEMS Tests were designed for upper primary and lower secondary school children, many of them provide diagnostic material for use with any group of learners. Attempts were also made with some of the tests to assess the learner's communicative competence independently of his linguistic competence. For the purpose of the tests, "communicative competence" was defined as "the speaker's ability to understand the social significance of utterances, and to produce utterances which appropriately reflect the social norms which govern behaviour in specific encounters" (ACER, 1976:82). The problem of defining and testing "communicative competence" is now more than ever a central one in any discussion of language proficiency.

In developing tests of language proficiency in the context of an oral interview the TEMS material is a valuable resource that is both comprehensive and underused. It can also be used as a model in terms of the scope and depth of language assessed. In the initial stages of this project, however, only a restricted range of objectives and test items can be developed leading to a smaller test battery. It is envisaged that in subsequent stages it will be necessary to expand the test battery to calibrate the items on a larger sample of students.

Discussion with teachers at the AMES also revealed work done in syllabus development. A Syllabus Guide (AMES, 1983) revealed a structural/functional/situational syllabus provided for the use of teachers. This is a hierarchically arranged grammatical syllabus in which:

> ... grammatical structures have been linked to functions so that the students' immediate needs are covered within a context of structural progression. (AMES, 1983:1)

A list of situations designed to exemplify the structures and functions is provided for the teacher's guidance. In addition, teachers working in various AMES centres around Melbourne had developed their own syllabi designed to co-ordinate teaching with the current practice of grouping students by the use of the ASLPR. In the syllabus proposed by Fortington and Cartwright (1985), a hierarchical progression of grammar/language elements has been proposed which teachers might exploit according to their own professional judgement. Thus a teacher might find that these elements could be best taught in a structural, functional, situational, process or behavioural approach or any adaptation of these, the teacher choosing the most appropriate for any given class. A greater emphasis on the functional/notional approach with the language elements attached is that developed at the Midway Centre. These syllabi accord with the eclectic approach to classroom practice observed at Myer House. The diversity of courses and syllabus materials that the teachers use are in accord with Clark's notion that a syllabus is:

> ... an expression of what is deemed relevant and possible to teach within a set limit of time to a particular group of learners who are aiming at mastery of a certain number of activities to acceptable levels and who in the successful performance of these activities will show enormous variation. (Clark, 1979)

Teacher Reports

The general approach taken by the teachers in the reports is exemplified in the following figures. Several points arising from them warrant further attention. The thrust of the reports is described under the headings "Objectives" and "Course Content". The teachers' interpretations of the end result of the course is described in the column headed "Learning Outcomes".

LAID Course Type Focus-Purpose Macro skills	Client Profile	Objectives	Course Content and Methodology	LEARNING OUTCOMES
LAID Survival Language On-Arrival Full-time ASLPR 1- Focus - Oracy Language in every-day situations Macro Skills S - 1- L - 1- R - 1- W - 1-	Students lacked necessary language structures, vocabulary and confidence in interaction situations, particulary in the listening/speaking skills. New arrivals Small, homogeneous young groups (Turkish (Vietnamese, Chinese (Croatian, Japanese The Asian student (4) had very poor pronunication Most of the students had little and/or confused knowledge of grammar structures, tenses etc All were initially unable to produce longer utterances.	To improve performance in relevant situations and theme areas: e.g. Personal I.D. Socialising Public Transport Employment, Health To create a relaxed conducive learning environment to overcome shyness and boost confidence. To improve the sounds they had difficulty with also rhythm, stess and intonation To teach and consolidate fill sound, the following tenses:Present, Pres-Cont Past, Past Cont., Present Perfect, Future and other useful structures occurring in context. To improve discourse skills and students' ability to produce complex sentences.	Core text: "Using the System" C Corbel. Developed a role-play and authentic practice situations as much as possible. - Group/Pair work - Excursions - Multi-Cultural morning teas. Special pronunciation tutorials, and regular pronunciation work prior to role-playing. Teacher presentations followed by related exercises form "Using the System" also "Side by Side" Sentence building activities, word order linking exercises. Functions:Giving Opinions Agreeing/dis-agreeing Discussion/Conversation	Students appeared to be able to use and transfer appropriate language to known and less familiar situations. Their general comprehension improved. Students worked well together helping and correcting each other. They showed interest in each others culture and out of class went to restaurants, gym and aerobics together. 3 of the 4 spoke more clearly by the end of the course, with a conscious effort to put tongue in correct position, e.g. R/L Students were using correct tenses with correct time phrases. If they made a mistake they's usually self-correct Students could answer open-ended questions fully. All students gave excellent multi-cultural talks and answered many questions. Most participated well in class discussions.

Figure 4: Sample Teacher Reports

LAID Course Type Focus-Purpose Macro skills	Client Profile	Objectives	Course Content and Methodology	LEARNING OUTCOMES
LAID On Arrival part-time ASLPR "0" Focus: Oracy language in situations Purposes: Survival	Nationalities 3 Timorese 1 Vietnamese 2 Chilean 1 Greek 1 Peruvian 2 Macedonian 1 Turkish Educational Background 5 - 16 years Age Range 18 - 43 Motivation varied from high (Greek dentist) to unaware Vietnamese youth- subsequently transferred) All very inconfident of any utterances 1 student left (2nd day) 3 found jobs.	Students to be able to use present tenses of common verbs to be to have to go correctly .....etc In order to be able to give personal information direction appointments interview etc. to be able to use past tense of, a few verbs with acceptabl precision	Routine use made of the following texts: "Streamline Dept" used for dialogue modelling with tape for listening exercise and re-production, for reading all in a structural "mode". "Tree of Three" used as complementary material focusing on Pronunciation related to structures learnt (above) Also Jazz Charts walk to work, Side by Side Contact Pictures, Video Using the System etc.	On Post Course Assessment for On Going courses of 7 students completing the courses 5 were graded "1-" ASLPR; the other two "0+". This was most significant improvement for an initial "0" class. Further, confidence was developed in ability to cope with everyday situations and communication in social environments in party practice other groups The student only felt the lack of more classess to go on to

It appears that the teachers use objectives to define the course. However, the interpretation and development of objectives are not consistent across teachers or across courses. This makes it difficult to use teachers' objectives as the starting point for the development of proficiency assessment or even for assessment of learning outcomes for specific courses.

An examination of Figure 4, however, illustrates the varying interpretation of the term "objective". The level 0 example objectives define the student level performance. The examples for levels 0+, 1- and 1 describe the teachers' intentions, and the example for level 2 has a mixture of these two interpretations.

The course reports provided an information base which enabled a further examination of emphasis on spoken language in courses offered by the AMES to be undertaken. In each course report, key terms relating to language development and instruction were identified and classified under seven major headings. It would be possible to identify more categories or to reduce the number, but the classification for the content analysis is based on the dicussions of language development and the elements deemed to be important in the literature on the subject.

(i) The CONTENT OR BACKGROUND refers to the first language, educational level of the clients, the nature of the course, existing knowledge of English, additional tuition and so on.

(ii) FLUENCY has been used to classify terms such as "speed", "fluency", and "fluent responses".

(iii) FUNCTION has been used as a classification to enable counts of terms such as "seeking employment", "giving opinions", "communicating", "conversation", "dialogue", "seeking and giving information", "asking questions" and "giving directions". There were obviously many different ways in which functions were referred to in the text of the reports.

(iv) STRUCTURES were obvious in that terms such as "syntax", "tenses", "verbs", "structures", "grammar", "complex sentences", and specific examples of tenses were grouped together under the broad heading of 'structures'.

Table 1 Percentage Frequency of Terms Used in Course Reports

	ASLPR Level						
	0	0+	1-	1/1+	2/2+	?	Mean
Context	10.3	9.3	2.1	9.1	3.6		7.1
Fluency	6.9	5.8	6.4	5.4	7.1	5.8	5.4
Function	13.8	18.6	14.9	13.6	21.4	19.8	17.0
Structure	41.4	32.6	29.8	37.9	32.1	31.5	34.7
Vocabulary	13.8	8.1	6.4	6.8	21.4	17.4	9.4
Pronunciation	3.4	15.1	29.8	12.9	3.6	18.6	14.2
Social	10.3	10.5	10.6	14.4	10.7	6.9	12.2
	100	100	100	100	100	100	100

? unlabelled level

(v) VOCABULARY was used to group terms such as "vocabulary", "register", "word use", "reproduction" and "lexis".

(vi) PRONUNCIATION was used to classify terms such as "rhythm", "stress", "pronunciation", "tongue placement" and "intonation".

(vii) SOCIAL/PERSONAL aspects were identified using terms such as "cohesion", "self-confidence", "motivation", "group binding", "social interaction", "cultural influences" and "personal needs".

These classifications were used to obtain a frequency count associated with each classification across all course reports provided. The results of the content analysis are presented in the table below.

At every level the most emphasised component is structure. Function, pronunciation and social aspects are also given considerable emphasis but not the same relative importance as structure.

A surprisingly low rate of mention was given to vocabulary at the lower ASLPR levels. This may be due to a number of factors: acquisition of vocabulary usually takes place by reference to the student's mother tongue, or at a more advanced level by explanation in the target language in which synonymous

expressions are used. Given the heterogeneous nationalities of the clients, the first possibility could not occur and, given their minimal English, the second would have been impossible.

During the classroom observations, however, some students were observed to be making extensive use of dictionaries. The vagaries of English spelling in this case could cause great frustration to the student unless the new vocabulary were listed on the whiteboard.

A further explanation could be seen perhaps in the methodological approach of the teaching, which for beginning students consisted of a form of the direct method based on a functional content where possible. The learning of formulaic phrases in this sense might not be seen by teachers as vocabulary acquisition but be considered as "given". Again it might be argued that the constraints of a direct method approach do not assist fast vocabulary acquisition.

At high levels, 2/2+, vocabulary and structure are given equal emphasis. It is also possible that vocabulary was never emphasised as such because it is built into specific functional and background aspects of the courses. If these areas are taken as broad bases for vocabulary development in the courses and vocabulary added to function and background to estimate the emphasis given to vocabulary, then the emphasis rises considerably. However, this is conjecture and the frequency with which teachers report on vocabulary development indicates that there is little emphasis given to this aspect in the AMES courses, or at least it is not emphasised at assessment stages.

Social and personal aspects of the courses were given a relatively consistent emphasis as was fluency. This is somewhat surprising given that fluency and social and personal aspects of language development may be difficult to develop at the lower language levels where basic language and vocabulary need to be developed. Again a possible explanation may be that the teachers take these aspects as givens and interface them into courses as a natural aspect of the functional background and social aspects of the course. Alternatively, social and personal aspects are treated as synonymous with vocabulary and simple structure at this time.

The implications of the analysis of the teachers' course reports seem clear. The emphasis in an instrument designed to measure achievement should reflect the importance given to structures, functions, pronunciation, vocabulary and social and personal factors. In developing measures of proficiency, it has already been argued that the instrument must be transferable across courses.

Accordingly, the relative emphasis should reflect the emphasis given across a wide range of courses and, as such, should reflect the emphasis given in the course reports across the assessed proficiency levels.

It is further evident that the teachers' reports cannot provide the sole basis for an overall approach to the development of generic objectives. The differing interpretation of the nature of an objective means that a format cannot be developed from the course reports. However, the general context of the objectives may be used. It is clear that they need to address structures, functions, context and possibly vocabulary, if the assumed role of vocabulary is as described above. Personal and social aspects do not appear to have as much emphasis required as the other elements at the levels of language focused on by the AMES.

The difficulty of accommodating both structure and function may also be guided by the analysis of the teachers' reports. Wilkins' two categories are also evident in the reports. It is clear that at low levels of language, the structural elements are considered by the teachers to be more important than functions, although this does not imply that functions are not important. Simple functions might be used to teach a large range of structures or, alternatively, learning a restricted range of functions may require implicit or even explicit instruction in appropriate structures to enable relevant material and skills to be covered within the possible limits of the students' language acquisition. The implications of the teachers' reports suggest the "ladder and ribbon" model of Brumfit (1981).

Classroom Observations

The data gathered in this section of the project included extensive observation of lessons in classes at different levels of language acquisition. Classroom observations were undertaken for three reasons. The first was to get to know the teachers and students and become more aware of their requirements. The second purpose was to examine the content of classroom instruction. This would enable us to establish what teachers and students thought were the important elements that led to an improvement in language proficiency and how they were translated into classroom practice. Finally, it was expected that classroom observation would be a valuable resource for the development of assessment techniques. It was hoped that key aspects of teachers' assessment methods could be translated into techniques for testing in a proficiency interview.

Placement into the observed classes had been determined by an oral interview based on the ASLPR. All relevant skills were taken into consideration. Discussions with teachers during this phase of the project as to the content of courses being taught revealed a desire to meet the students' "needs". These were expressed in a variety of ways ranging from "building up the students' confidence", "they need more listening", "they can write it but they can't say it" (a reference to the pronunciation problem of the morpheme "s", characteristic of Vietnamese students), to discrete point lists of the students' language needs. (The students need to find out and use relevant language for health, housing, employment, personal identification, orientation, services - including shopping, safety, the law, the media, letter writing, discussing social customs and so on.) These categories could be largely subsumed in the discussion of learner language "needs" as indicated by Clark (1979). Further discussions with teachers across classes were expected to determine commonality of the underlying or generic nature of discrete courses and the teachers' methods of intuitive or progressive assessment. However, verbal assessment in the classroom consisted mainly of recitation exercises and a verbal equivalent of the cloze or sentence completion procedures. In all approaches to classroom monitoring, the student was prompted or even given the answer in the eliciting language. It was doubtful whether any true assessment occurred in these exercises (as distinct from teaching and assessment combined) and there did not seem to be a great deal of purpose in repeating these techniques in a formalised proficiency test. The difficulty of separating teaching and testing proficiency did not appear to have been overcome. This conclusion was reinforced during assessment workshops organised for teachers as part of this project.

The concentration on teaching is understandable but the assessment techniques would not be likely to lead to reliable or valid information about achievement or proficiency. Only intuitive judgements could be made, largely based on the experienced teacher. For the less experienced teacher there did not seem to be any systematic way of monitoring achievement or developing proficiency and this made the end-of-course interview all the more important. Because of the diverse nature of the courses, their content and selection of a range of functions, the only substantial data available on development was an end-of-course ASLPR classification. This made further exploration of proficiency rather than achievement, an indispensable focus of this project.

AN ACTUARIAL APPROACH

Researchers in Australia (Pienemann, 1983; Johnston, 1985a; Johnston, 1985b) working in the area of language acquisition on an actuarial basis have produced a data map for a cross-sectional study of language acquisition. The data was collected by means of unstructured conversations and then analysed for the emergence of certain morpho-syntactic structures and information on lexis. The project restricted analysis to the area of morpho-syntax, and only a small sample of Vietnamese and Polish immigrants was used. Nevertheless, the results appear to offer an acceptable and viable grammatical model which does not conflict with that proposed by others working from an analytical approach.

Whilst the amount of information provided by Johnston is huge in respect to quantity, the sample size is small. The stability of the data may therefore be low. However, certain conclusions may be drawn about cross-sectional differences in language proficiency.

Johnston offers a limited amount of data describing longitudinal development of proficiency. While it is limited, it is rare that a research study even aims to study development within individuals over time. Despite the paucity of stable cross-sectional data and the small amount of developmental data it is nevertheless a rich resource in that definite patterns of language use have been identified across levels of differing proficiency and by individuals across time as proficiency develops. The end result of Johnston's work is that it becomes possible to draw tentative conclusions about proficiency development.

These are expressed in terms of the syntactic acquisition of language and they indicate that second language learners tend to produce language systematically. Systematising appears to be related to the learner's mother tongue proficiency. Different learners will show individual variations but a continuum of increasing proficiency could be shown to exist. It is described as an "implicational relationship" and enables a prediction to be made about what a particular learner can or cannot be expected to do. Thus the data would suggest a hierarchy of acquisition: a learner who uses "could", for example, will almost certainly use "must" and "can". The acquisition sequence of modals can be said to constitute a developmental sequence. This developmental sequence as described by Johnston bears a remarkable similarity to the notion of the measurement model described in the introductory sections of this paper. Other sequences of acquisition, i.e. for prepositions and

irregular verbs, also occur, but it is not yet shown how the sequences of acquisition are related to one another or even if they are. Work in this area is currently being undertaken. Assistance with an explanation of these problems has been sought in speech-processing theory (Pienemann, 1983). As speech events occur in "real time" and in many cases involve many processing subroutines, the conscious mind can only focus on a limited part of the whole speech-processing operation. Until certain constituents of a speech act are automatised, a trade-off situation will exist so that a fully correct response to a stimulus will not be produced, depending on the complexity of the speech processing required. As a result of this input from speech-processing theory, features of learner language not constrained by developing speech-processing mechanisms were termed "variational", and were seen to correlate with the learner's "psychological" situation, varying according to the learner's mental make-up and position and prospects in the host society (Johnston, 1985b). This variational influence is described by Johnston as a second dimension. The implicational dimension however has been described in a series of six stages, through which a student passes. At each stage the level of competence in a range of grammatical areas becomes greater. The major grammatical areas or elements which Johnston (1985a) has identified are collected together in his grammatical model. These elements are listed below.

1) Verb morphology - "-ing", "-ed", "-s" marking, and so on.

2) The use of the verb "to be".

3) The use of modals - "can", "will", "must", etc.

4) The development of negation - use of "don't", "any", etc.

5) The use of questions.

6) Noun morphology - the use of plural and possessive "-( )s", etc.

7) The use of definite and indefinite articles.

8) The use of quantifiers - words like "all", "every", "some", etc., and also the use of numbers themselves.

9) Deixis - the use of words like "this", "that", "here" and "there", as locators of events and things in space, time and discourse.

STAGE	VERB	NOUN	PN	Q	NEG	ADV	ADJ	PREP	W_ORDER
1:	'WORDS'	or	FORMULAE						
2:	•	•	1st	SVO?	no	-	-	PP	SVO
	IL-ing	•	2nd	•	no + X	-	-	•	•
	IRREG	•	3rd	•	•	-	-	•	•
3:	-ed	REG_PL	POSSESS	DO_FRONT	don + X	(ADV)	-	•	TOPIC
	•	IRREG_PL		WHX_FRONT	• .	-	(more)	•	ADV_FRONT
4:	AUX_EN	(POSSESS)	•	PSEUDO_INV	•	•	(better)	COMP_TO	PART_MOV
	AUX_ING	•	•	Y/N_INV	•	•	(best)	•	PREP_STRNDG
5:	3SG_S	(PL_CONCD)	CASE(3rd)	AUX_2ND	DO_2ND	-ly	-er	•	(DAT_TO)
	•	•	RFLX(ADV)	SUPPLET	SUPPLET	•	-est	•	•
6:	(GERUND)	•	RFLX(PN)	Q_TAG	•	•	•	•	ADV_VP
	•	•	•	•	•	•	•	•	(DAT_MVMT)
	•	•	•	•	•	•	•	•	(CAUSATIVE)
	•	•	•	•	•	•	•	•	(2_SUB_COMP)

KEY: (Round brackets indicate tentative assignment only).

IL-ing = non-standard"ing"; PP = in prepositional phrase.
DO_FRONT = yes/no questions with initial "do"
WHX_FRONT = fronting of wh-word and possible cliticized element (e.g. "what do").
TOPIC = topicalization of initial or final elements;
ADV_FRONT = fronting of final adverbs of adverbial PPs.
AUX_EN = (be/have) + V-ed, not necessarily with standard semantics.
PSUEDO_INV = simple fronting of wh-word across verb (e.g. "where is the summer?")
COMP_TO = insertion of "to" as a complementizer as in "want to go".
PART_MOV = verb-particle separation, as in "turn the light on".
AUX_ING = (be) + V-ing, not necessarily with standard semantics.
Y/N_INV = yes/no questions with subject-verb/aux inversion.
PREP_STRNDG = stranding of prepositions in relative clauses.
3SG_S = third person singular "-s" marking.
PL_CONCD = plural marking of NP after number of quantifier (e.g. "many factories").
CASE(3rd) = case marking of third person singular pronouns.
AUX_2ND = placement of "do" or "have" in second position;
DO_2ND = as above, in negation.
SUPPLET = suppletion of "some" into "any" in the scope of negation.
DAT_TO = indirect object marking with "to".
RFLX(ADV) = adverbial or emphatic usages of reflexive pronouns.
RFLX(PN) = true reflexivization.
Q_TAG = question tags.
DAT_MVMT = dative movement (e.g. "I gave John a gift").
CAUSATIVE = structures with "make" and "let".
2_SUB_COMP = different subject complements with verbs like "want".

1. Source: Pienemann and Johnston (1985).

Figure 5: Johnston's Model of Language Acquisition

10) The use of existential propositions (equivalents for sentences involving "there is/are" in standard English).

11) The use of personal pronouns.

12) The use of prepositions.

13) The use of connectors - words like "and", "but", and "if".

14) The development of vocabulary.

Figure 5 above illustrates how these grammatical elements develop over the stages of proficiency.

This model developed by Johnston appears to be particularly promising but at this stage it is not sufficiently developed to form a basis for a set of objectives. The greatest contribution of the work of Johnston is the implications that it has for the development of response criteria for the objectives.

In following the approach of Johnston and Pienemann, scoring criteria for the objectives were established after the objectives were trialled. During a workshop with teachers, students' responses to sample questions were closely examined. By listening to the differences between the responses of students at a range of proficiency levels, as defined by the ASLPR, the contrasting features of response could be used to define criteria.

CHAPTER 6

OBJECTIVE CONSTRUCTION

Objective Style and Organisation

The style of objectives originally proposed for the AMEP followed that of Mager (1973). This form of objective requires a precise description of task and standards. The tight specification of the tasks would have required an enormous bank of objectives considering the observed differences in classroom methodologies and the variety of resources, techniques and contexts used in courses. Further, the tight specification of performance criteria in the objectives would have led to discrepancy-type evaluation that has an in-built notion of failure. The objectives as prepared in the AMES proforma may have benefit in assisting teachers to plan instruction, and may be generated from a generic style of objective, discussed more fully below.

However, we regarded this approach to test construction as unsuitable for two major reasons. First, the style of the objectives restricted their application to the specific course for which they were designed. Secondly, the heterogeneity of format and style of such a set of objectives, along with the discrepancy approach to evaluation, left the research team uneasy with the assessment style. It was clear that the objectives needed to be provided with some form of organisation so that progress could be monitored. However, this implies that there are paths along which a development can be traced. The literature refers to these paths as dimensions, but there appears to be no resolution regarding the nature or the number of these dimensions. It was decided, therefore, to adopt a dimension-based model - namely the Partial Credit Model - including the accompanying assumptions rather than have a disjointed batch of objectives and a discrepancy approach. This enabled a developmental rather than discrepancy basis to be used in test construction and offered the chance to examine the progess and success of clients rather than their failure or discrepancy from pre-specified discrete point standards.

There are many theories regarding language development and a large number of these theories have their merit in explaining aspects in an individual's changing ability in a second language. There is no lack of theory development, but little attention seemed to be paid to sound theory testing.

In factor analytic studies used to demonstrate dimensionality, it is common for a 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. 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. Therefore, the issue of dimensionality of 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 programs round each, completely isolated and unrelated to programs 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 contained. These single or multiple factors are in fact manufactured by the analytical methods and the measurements used (Carroll, 1983; Bachman and Palmer, 1983).

Studies that are not factor based put forward theories of language development and proficiency which are based on very small samples. 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.

In summary, studies based on very small samples tend to yield broad generalisations beyond their external validity, or a quantitative approach is adopted and 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 specification of the model is given little or no attention and it is tested with measures of unknown measurement properties. It is even possible that the simple or sophisticated statistical analyses are conducted oblivious to the

fact that each analysis assumes a mathematical model and that this assumption implies that the researcher believes that language development and differences among individuals can be summarised in a mathematical equation. Many of the results can be dismissed in large measure simply because the mathematical model underlying the statistical analysis would be considered outrageous.

Organising Framework

Rather than taking a set of measures of unknown properties, the logical way to investigate a dimension is to determine if one can be defined and then test it with a model designed specifically for the purpose. Such a mathematical model is the Partial Credit Model. This first investigation was begun with a dimension that could loosely be termed grammatical competence. This organisation began with the Relative Contribution Model proposed by Higgs and Clifford (1982). Essentially the dimension defined begins with isolated elements of vocabulary. It then moves into the use of some basic formulaic language and basic structures followed by the more di ficult grammatical elements. At a later stage it was possible to map this dimension in more detail through the aid of the data collected, and hence show that, it was difficult to argue that the dimension does not exist.

Numerous definitions and discussions regarding the dimensionality of language proficiency and communicative competence exist in the literature (e.g. Canale and Swain, 1980; Oller, 1983; Hughes and Porter, 1983; Higgs, 1984; James, 1985; Rivera, 1985). This debate is not under discussion here and the dimension in this study is chosen not because it is the only dimension but because it is an important possible dimension. Evidence from the available literature and data gathered from classroom observations and discussion was regarded as important in the development of adults' English.

As described earlier, a check on the appropriateness of applying the Higgs and Clifford model to local conditions was undertaken at two workshops with AMES teachers. At these workshops teachers were also asked to order objectives developed from the model accordingly and place them at an ASLPR level. The objectives in the Testing Manual are ordered according to the original teacher rankings.

The Objectives

Each objective in the Testing Manual is specified by six separate pieces of information. The General Objective specifies the focus of the objective in a functional type form. The Possible Language element specifies the most likely language element or structure that will be used by the client in response to a question or in performing a task based on the objective. It is the language element that has been used to provide the organisation for the objectives. However, a client's response to an item developed to test an objective may be regarded as completely appropriate even when the possible language element is not contained in a response. This measure, therefore, provides flexibility for creative uses of language which have not been foreseen.

The Question section contains three subheadings: Type Formula and Samples. Type is a brief description of the type of question that should be constructed. The formula gives the precise form of the questions that can be written to test the objective. The samples are a number of sample test questions that have been written to test the objective. They show some possible items that can be written to satisfy the formula specified and test the objective.

The Restrictions and Instructions cover any further information that is required when developing a test item for the objective and administering it to a client. The restrictions provide information that is required in writing test items. They provide general constraints on the vocabulary and contexts that can be used in developing the items. The instructions provide information that should be used by the interviewer when administering the test items. The Response Criteria specify the criteria required to score an item and are to some extent specific to the item used to test the objective. In the objectives the response criteria correspond to the items used in the trial testing and test development described in the next chapter. In general they can be easily modified to suit the context of the syllabus and the stimulus that is being used to test the objective. In this modification the focus of the response criteria must be maintained.

Constructin, a Test Item

The formulae provided with each objective provide a framework for constructing an item bank of test questions that have the same focus but may be applied in different contexts. The formulae should not be seen as an attempt to define a set of syntactic or grammatical rules for the development of language. The

formulae are couched in a loosely grammatical form to ensure that only a restricted and specific range of language is used by the interviewer when developing items to test each objective. The symbols in the formulae are defined as follows:

< >	Elements in angular brackets are replaced by a word or phrase. Often restrictions on the possible inclusions are specified.
| |	Elements listed in between straight lines are alternatives. One of which must be used.
[ ]	Elements in square brackets are optional, i.e. they may be omitted.

By using the formulae each objective can be tested by an item that is developed to suit the content required by the teacher. For example objective 11 has the following formulae:

<Pronoun> / <noun> | want/s [<verb>] | <noun subject> / <pronoun> | [<verb>] <modifier> <noun>.

What | does / do | <noun subject> / <pronoun> | say / ask | [<pronoun>]?

This could be tested using any of the following questions:

1. Lyn wants her friend to come to her home. What does she ask her?
2. You want your friend to have a drink. What do you ask him?
3. She wants to buy her friend a drink. What does she say?

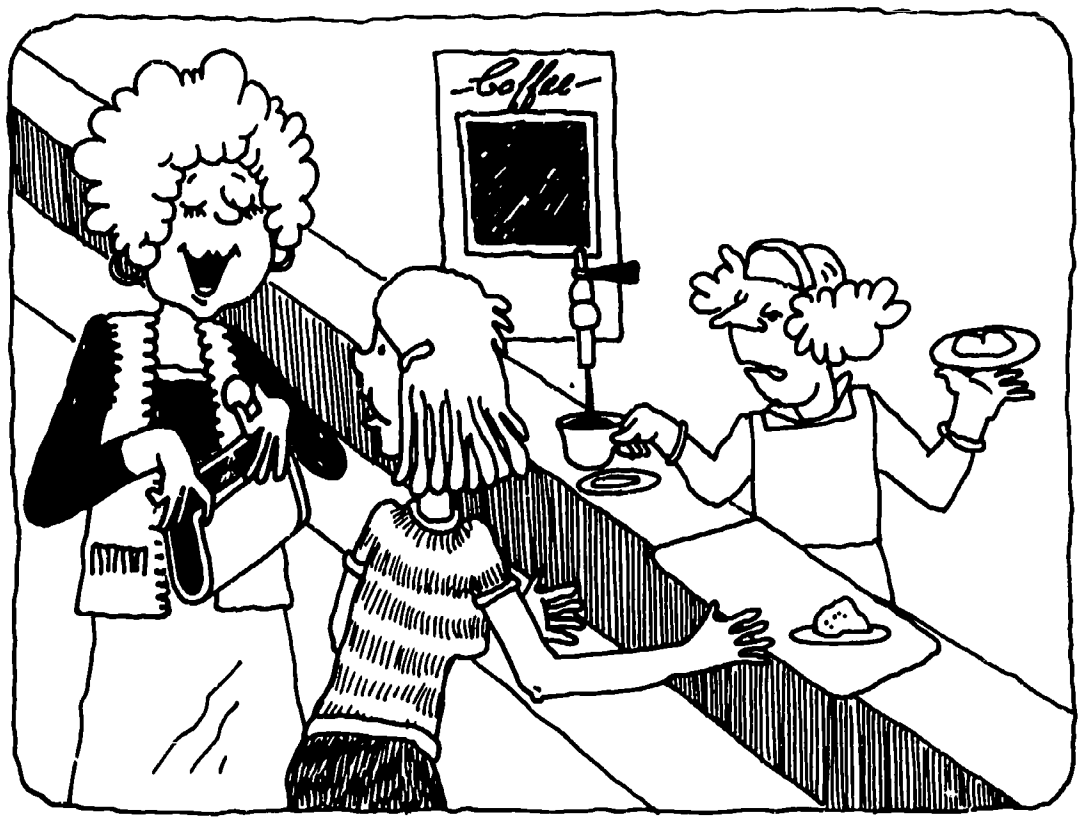


Figure 6 An Item to Test Objective 11

The example item used to test this objective in trialling is shown in Figure 6. The interviewer used this stimulus with the prompt: "She wants to buy her friend a coffee. What does she say?"

Each of these examples tests the same objectives but may or may not be suitable for testing achievement after a particular course. Provided the formulae restrictions are followed it is possible to create interview test questions that are suitable for a wide range of possible contexts. Further, by strictly adhering to the formula approach, it becomes possible to avoid giving the appropriate response embedded in the stimulus, thereby encouraging an "echo" response. Each objective has been designed to avoid both the echo and the verbal equivalent of a cloze test item.

CHAPTER 7

DEVELOPING AN EXAMPLE PROFICIENCY TEST

To obtain an adequate data base to undertake these analyses, an oral interview based test was developed and administered to 270 students classified as having proficiency levels from 0 to 1+ on the ASLPR. A subset of twenty-nine of the objectives were then developed into test items. The organisation of the items into three testing forms is shown in Figure 7. For the trial testing the items were divided into four subsets of six to eight items each and then grouped to form three tests. The items in subset 2 were used on both tests A and B and the subset 3 items were used on tests B and C.

The test was not scored during the trial administration. Each of the interviews was recorded and scored at a later date according to scoring criteria developed in workshops with teachers. To develop the criteria the responses of students at a range of proficiency levels were examined and characteristics of the responses were identified. Scoring criteria were then constructed to distinguish between students. The tests were then scored from tapes by nine raters, including members of the research team and AMES teachers.

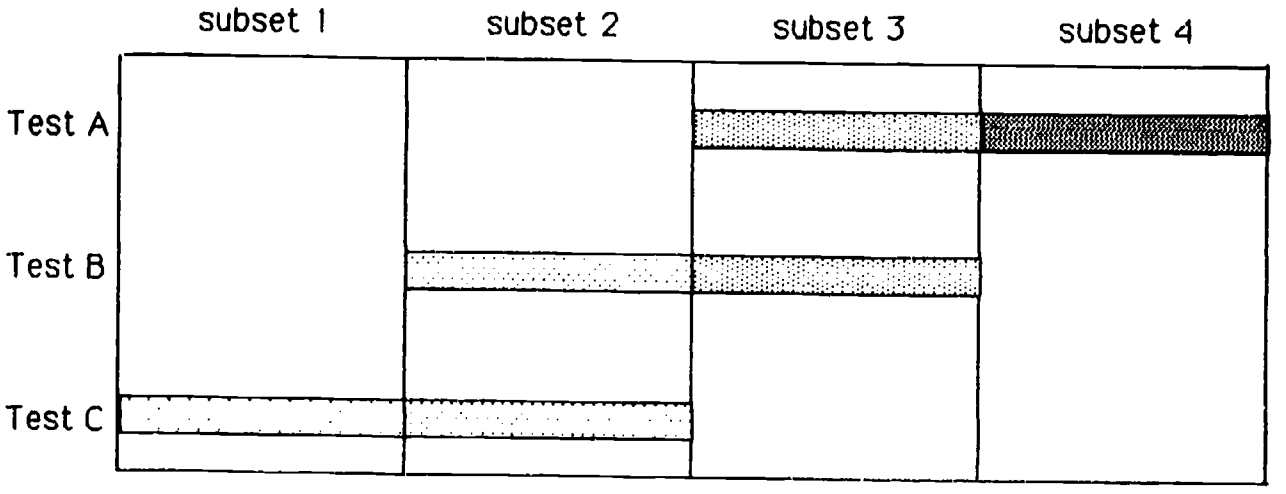


Figure 7 Structure of the Trial Testing

Test Analysis

After collecting and coding the data each test was analysed using the CREDIT computer program (Masters, Wright and Ludlow, 1980). This program applies the model described in Chapter 4. Tables 2, 3 and 4 below show the results of these analyses, including the partial credit analyses and descriptive statistics for each test.

Table 2 Item and Test Statistics for Test A

Test A	d(i,1)	d(i,2)	se(i,1)	se(i,2)	fit
1.1	-2.62	-2.01	.76	.37	.69
1.2	-1.15	- .75	.38	.31	2.36
1.3	- .83	-3.64	.67	.53	- .18
1.4	-1.05	- .69	.37	.31	.41
1.5	.27	2.42	.30	.62	- .13
1.6	-1.23	.55	.34	.33	.13
1.7	-1.32	- .35	.38	.31	- .68
1.8	-1.88	1.58	.38	.39	- .83
2.1	- .34	1.29	.30	.40	- .08
2.2	- .15	.80	.31	.37	.96
2.3	.53	1.30	.31	.45	1.15
2.4	.22	1.08	.31	.41	-1.77
2.5a	- .26	3.02	.29	.73	-1.86
2.5b	- .07	2.53	.30	.61	- .28
2.6	.36	2.39	.30	.62	- .52

N = 58 $\bar{X}$ = 14.22 α = .80

Table 3 Item and Test Statistics for Test B

Test B	d(i,1)	d(i,2)	se(i,1)	se(i,2)	fit
2.1	- .86	.71	.28	.32	- .01
2.2	-2.22	.10	.37	.27	- .85
2.3	- .33	- .08	.28	.29	1.17
2.4	-1.59	.69	.31	.30	1.03
2.5a	-2.21	1.84	.34	.40	- .10
2.5b	-1.57	.60	.31	.30	.29
2.6	-1.81	.90	.32	.31	.65
3.1	.57	1.11	.28	.41	.90
3.2	-1.14	2.49	.28	.53	-1.29
3.3	- .47	1.70	.26	.43	-1.56
3.4	-1.12	.17	.30	.29	- .15
3.5	-1.82	1.96	.31	.42	- .26
3.6	.19	1.65	.27	.46	- .17
3.7	-2.03	.36	.35	.28	.22
3.8	- .48	2.66	.26	.61	- .62

N = 74 $\bar{X}$ = 12.92 α = .81

Table 4 Item and Test Statistics for Test C

Test C	d(i,1)	d(i,2)	se(i,1)	se(i,2)	fit
3.1	1.08	.11	.30	.35	1.99
3.2	-1.98	1.58	.36	.35	- .71
3.3	.03	.68	.28	.34	.19
3.4	-1.38	.17	.34	.28	.93
3.5	-2.20	1.18	.39	.32	- .56
3.6	- .43	.75	.28	.32	-1.08
3.7	-2.41	- .23	.47	.27	- .30
3.8	-2.38	1.50	.41	.34	- .73
4.1	-1.74	.69	.35	.30	.52
4.2	-1.03	.64	.30	.30	1.25
4.4	-1.43	2.41	.31	.46	-1.59
4.5	.50	.57	.28	.35	1.16
4.6	1.83	1.82	.36	.39	1.03
4.7	.94	.62	.30	.30	-1.99

= 71 $\bar{X}$ = 12.88 α = .82

Each item in Tables 2, 3 and 4 is reported with two difficulty parameters, their corresponding standard errors 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. The fit statistic for items that conform to the model has an expected value of about zero and standard deviation of one. When the fit statistic exceeds two or is less than negative two ther is some doubt about whether this item works in the same way as the other items in the test. Previous research has shown that positive fit in excess of 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 in excess of two normally occurs when an item discriminates more highly than other items in the test. Tables 2, 3 and 4 also contain the mean and the standard deviation for each test along with the co-efficient alpha reliability.

One question (1.2) was found not to fit the model but examination of the misfit found that cause was scorer error rather than a flaw in the item. The strong fit of the data to the model and high test reliability clearly supports the hypothesis that the dimension as defined exists among the AMES students.

Of the twenty-nine items trialled twenty-four were retained to be included in the final proficiency test. The five rejected items were excluded because of problems with the art work and scoring criteria. They were not rejected due to lack of fit to the hypothesised dimension. The problems associated with the scoring criteria for items of this type are discussed in more detail in Griffin, Adams, Martin and Tomlinson (1985).

Equating Tests

Since the items on each test were calibrated separately the different tests needed to be equated before comparisons between items in different tests could be made.

In the trial testing the four item subsets were grouped to form three tests each consisting of two of the four item subsets. Subsets one and two were combined to form test A, subsets two and three were combined to form test B, and subsets three and four formed test C. In this design each person who completed a full test was therefore administered two of the item subsets.

The design of the trial testing, as shown in Figure 6 and described above, allows the items on the different tests to be equated through the application of common item equating. To perform this equating each test is calibrated separately, as shown in Tables 2, 3 and 4. Because the items in subsets 2 and 3 were placed on two of the tests, each of their items has two sets of difficulty parameters.

These difficulties can be used to perform common item equating between the three tests. The differences in difficulties for the items from subset two give the relative difficulty of tests A and B and the subset three items give the relative difficulty of tests B and C. In this case the shift required to equate the test forms is given by:

$$t_{AB} = d_{.A} - d_{.B}$$

where $d_{.A}$ is the average difficulty of the subset 2 items on test A and $d_{.B}$ is the average difficulty of those items on test Y. A shift t_{BC} can be similarly calculated. To estimate the quality of the link a number of procedures have been developed (Wright and Masters, 1982; Masters, 1984). Two of these procedures will be discussed as they are applied.

In Figure 8 and 9 plots are shown that compare the difficulties of the items on the two tests. In Figure 8, the difficulties of the items from subset 2 are plotted, and Figure 9 shows the difficulties of subset 3. For each plot a line of slope one, passing through the group means has been drawn. This line indicates the difference in difficulty for each item subset. The intercepts 1.307 and 0.490 respectively indicate the shift required to adjust the item difficulties to form a single scale. As expected in test construction the three tests are in ascending order of difficulty.

Since the measurements contain error, points plotted in Figures 8 and 9 are not expected to lie on the line through the means. To check the quality of the link the observed spread of the points around the plotted line are compared to what the model expects this spread to be.

If the data were to conform precisely to the model the variance of the points about the lines would correspond to the modelled variance given under each of the figures. The observed variance recorded under each figure is the amount by which these data were observed to vary about the line. The ratio of the

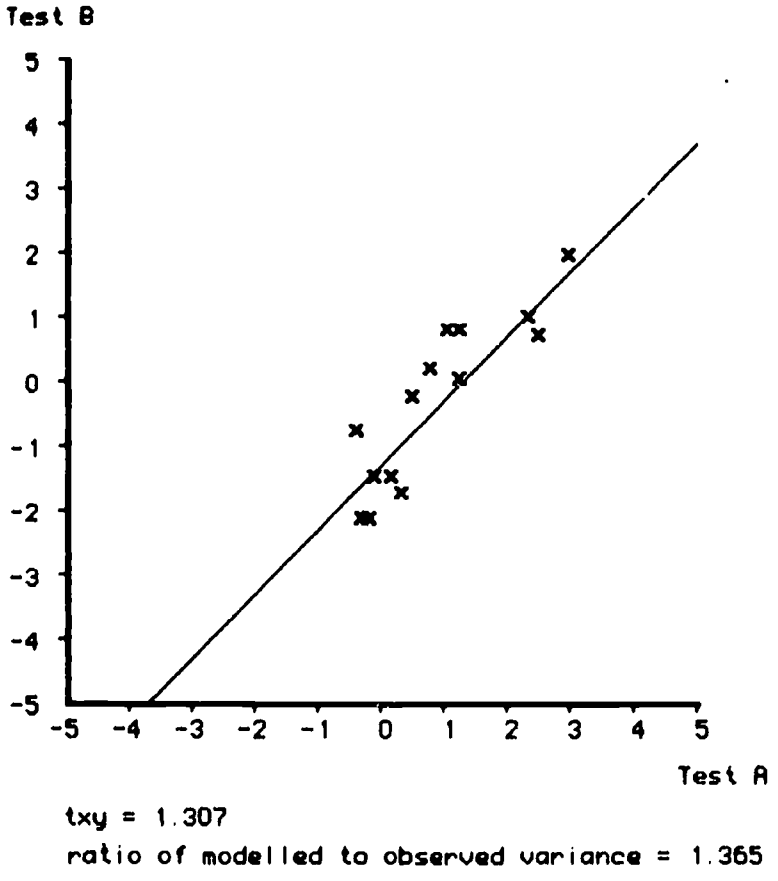


Figure 8 Plot of Subset 2 Items on Test A and B

modelled to the observed variance provides an indication of the fit of the data to the model and the quality of the links between items and subsets. In this case the ratios of 1.365 and 1.516 are considered very satisfactory. Ideally, this ratio should equal one, but in most studies reported in the literature a value of about 1.5 appears all that can be expected.

A second method to examine the link is to calculate a standardised difference for each pair of estimates after adjusting for differences in subset difficulty. That is:

$$z_n = \frac{d_{ijx} - d_{ijy}}{(s^2_{ijx} + s^2_{ijy})^{\frac{1}{2}}}$$

where S_{ijx} and S_{ijy} are the respective standard errors. An inspection of these values for the links indicates only three values outside the range -2 to 2.

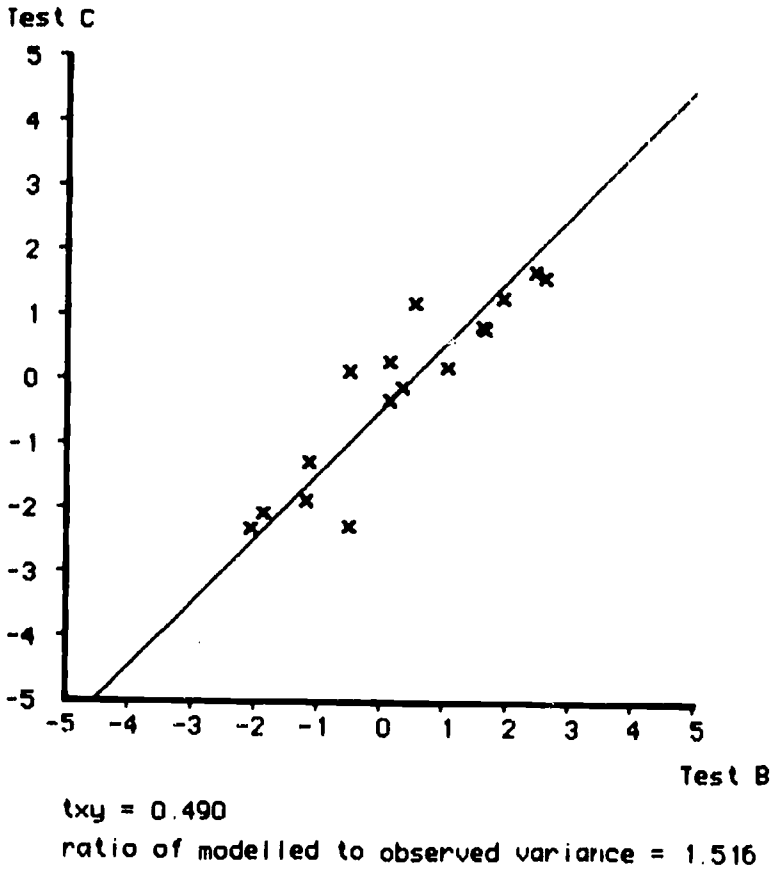


Figure 9 Plot of Subset 3 Items on Test B and Test C

The strength of the common item links supports the unidimensionality of the lack of test items. The links reported suggest that the tests are measuring on the same dimensions but at different difficulty levels.

Using the shifts reported above the item difficulties can be transformed to a common scale. In Table 5 the twenty items to be included in the central part of the final form of the test have been selected and their difficulties transformed to a scale ranging from about 10 to 90 (see Wright and Stone, 1980). This scale has been called the ITESL scale. Four other items are included in the ITESL test but they have not been equated into the ITESL scales. For more details of the test see the accompanying test manual.

Table 5 Equated Item Difficulties on the ITESL Scale

ITESL Name	Trial Name	d(i,1)	d(i,2)
1	1.1	14	24
2	1.3	20	17
3	1.2	23	38
4	1.7	30	38
5	1.6	31	47
6	1.4	32	35
7	3.7	35	57
8	2.2	36	54
9	2.5b	42	62
10	4.1	42	66
11	3.4	42	59
12	3.4	43	59
13	2.1	43	60
14	1.5	45	61
15	4.2	47	66
16	2.3	49	65
17	4.4	50	65
18	3 ?	53	69
19	3.6	54	69
20	4.5	63	66

Validity of the Test

In developing this test, construct and concurrent validity were of most concern. For our purposes construct validity refers to the existence of a measurable grammatical dimension in the test. During test and objective development a great deal of time was spent observing classes, discussing content, method and assessment with teachers and examining the literature on second language proficiency. On the basis of the data gathered in this process it was felt that there was sufficient evidence to suggest that a developmental dimension of grammatical competence exists in ESL and was of major interest in the teaching of adult migrants. A set of objectives was then developed to measure on this hypothesised dimension and a set of test items was created from the objectives. The existence of this dimension was then explicitly tested with the application of the partial credit model. In the application of the model we were unable to reject the existence of the

hypothesised dimension of grammatical competence as measured by the items in the test. It would appear from the analysis, coupled with the nature and sequence of the items, that the test items do work together to measure proficiency on a dimension so constructed. Further evidence regarding the validity of the test was provided by the successful linking of item subsets at different levels of difficulty.

We were also concerned with the tests' concurrent validity with the ASLPR. During the original data collection an estimate of the students' ASLPR rating was made by the interviewing teacher. After the calibration and scoring of the t[illegible] the correlation between ability as measured by the test and ASLPR scales was found to be 0.67. A plot of the students' ITESL and ASLPR scores is shown in Figure 10 below.

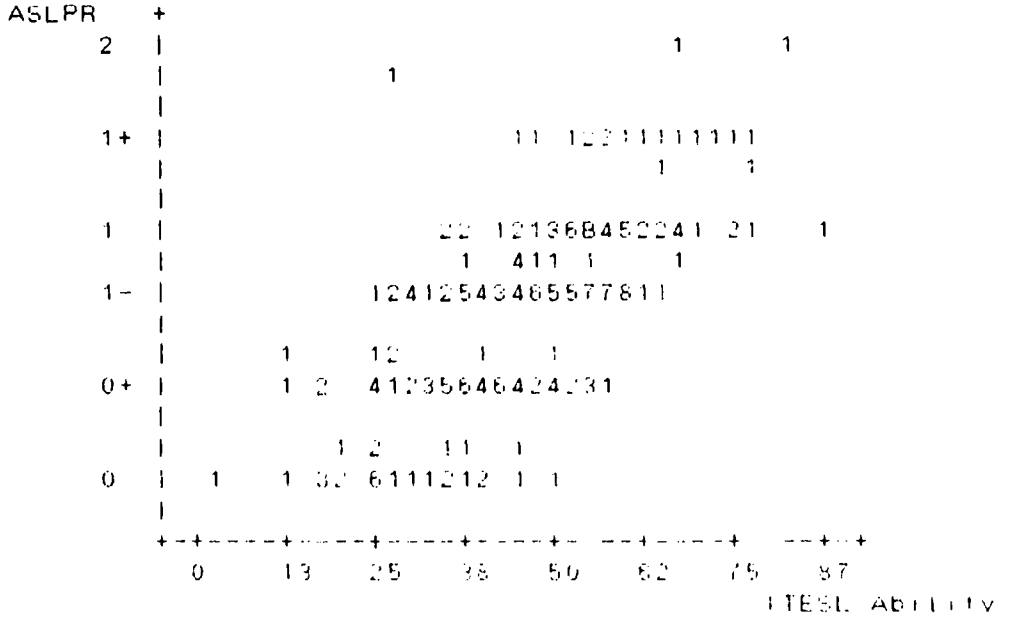


Figure 10 Plot of the ASLPR and ESL Scales

This is a particularly satisfactory result given that the ASLPR was scored on a six-point scale only and correlations with such a restricted range would normally be expected to be low. Secondly, the ITESL and ASLPR scales are not designed to measure identical constructs and, although it was hoped that they would be reasonably strongly related, a correlation that was too high would indicate that the ITESL and ASLPR scales were measuring exactly the same thing. If the correlation were too low we would be concerned that the two interview schedules were unrelated and that the ITESL scale would be of limited use. The correlation obtained was optimal.

Uses of the Test

The twenty items in Table 5 and the four additional items have been grouped to form three tests of ten items each and four general questions. Instructions for using and scoring the test are contained in the testing manual for the Interview Test of English as a Second Language (ITESL).

The major purpose of the Interview Test is to identify strengths and weaknesses in the client's English usage and to make decisions about appropriate instruction. The Interview procedures may be used for:

(i) placement of a new client;
(ii) monitoring progress during a course;
(iii) determining the outcome level of proficiency at the end of a course.
(iv) measuring achievement at the end of the course;
(v) diagnosing students' strengths and weaknesses.

Detailed instructions on how to administer the test are contained in the accompanying testing manual. In summary a student is administered one or two general questions followed by one test of 10 items. Test 1 consists of the first 10 items in Table 5, Test 2 consists of items 6-15 and Test 3 is items 11-20.

After administering the test the equivalence scales (Appendix B) are used to calculate a student's ITESL and ASLPR scores.

The ASLPR classifications have been included on the equivalence tables to provide a frame of reference within which to interpret the scores. The use of the ITESL score is recommended in preference to the ASLPR because of the finer gradations that it contains. In future testing sessions with the same student this allows for identifying improvements in students who might still remain within one classification on the ASLPR scale.

Each of equivalence tables provides four major pieces of information:

1. an ASLPR scale with regions marked showing the five lower ASLPR levels;
2. an ITESL scale marked in units from 10 to 90;
3. a series of marks corresponding to the possible scores on the test;
4. a plot of the item difficulties.

The scores 10 through to 90 are arbitrary and do not mean anything in themselves. They can be converted to any range of scores or descriptions the teacher wishes. However, the score range in the manual is recommended if only for consistent interpretation.

To read the equivalence table it will be necessary to use the ASLPR scale, the ITESL scale and the series of points corresponding to the student's score. The plot of item difficulties is explained in section 3.

A student's ITESL score can be calculated as follows:

1. Add up the student's score on the test.

2. Place a ruler horizontally through the point corresponding to the score.

3. Read the student's ITESL score at the point where the ruler crosses the ITESL line, and read the student's ASLPR from where the ruler crosses the ASLPR line.

For example, if a student was given Test 1 and scored a total of 10 then his/her ability on the ITESL scale would be 36 and his/her ASLPR 0+.

Monitoring Progress

By recording the student's ITESL and ASLPR score it is possible to monitor a student's proficiency development over time. Since the ITESL and ASLPR scales are independent of the items selected to test the student, it is possible to retest the student at a later date, possibly with a different set of items and then find the appropriate score on the ITESL and ASLPR scales. When using the test for these purposes care must be taken with interpretation of the score.

In Appendix B equivalence tables list the raw score on each test along with the corresponding ITESL score and standard errors. The standard errors can be used to establish performance regions by taking the ITESL score corresponding to a student's raw score and then adding and subtracting one standard error. For example a score of 10 on Test 1 defines the region from (36-5) 31 to (36+5) 41. On two separate occasions two regions can be defined and the smaller the overlap between the regions the more confident we can be that substantial improvement in performance has occurred.

The Item Plots

One of the uses of the ITESL is seen as the examination of the particular strengths and weaknesses of an individual student. To use the ITESL for diagnosis the plots of the item difficulties in Appendix B are used. These plots indicate what are called regions of most probable response. These regions indicate the expected response that a student with a given ITESL score will make to each item ("expected" here is not used in the statistical sense). Consider the item that is used to test nouns in Test 1. If a student was administered Test 1 and scored a one or a two then a ruler placed through the raw score of two as described in section 1 and shown in Figure 11 would indicate an ASLPR of 0 and an ITESL score of about 12. The ruler would also pass below the shaded region for the items labelled nouns. The expected response on the noun item for an ITESL score of below about 15 is zero. This indicates the student is unlikely to have a mastery of vocabulary relevant to the stimulus. If the student scored 3, 4 or 5 on Test 1, then the ruler placed on the raw score would pass through the shaded region for the noun item. This indicates that the expected response for the student on the noun item would be one, and that the student has, most likely, only a partial mastery of the relevant vocabulary. Similarly a client with a raw score of 6 or more on Test 1 has an expected response of 2 for the noun items. This is because a ruler placed on raw score of 6 or more passes above the shaded region for the noun item, and indicates to the test administrator that problems with the relevant vocabulary are unlikely.

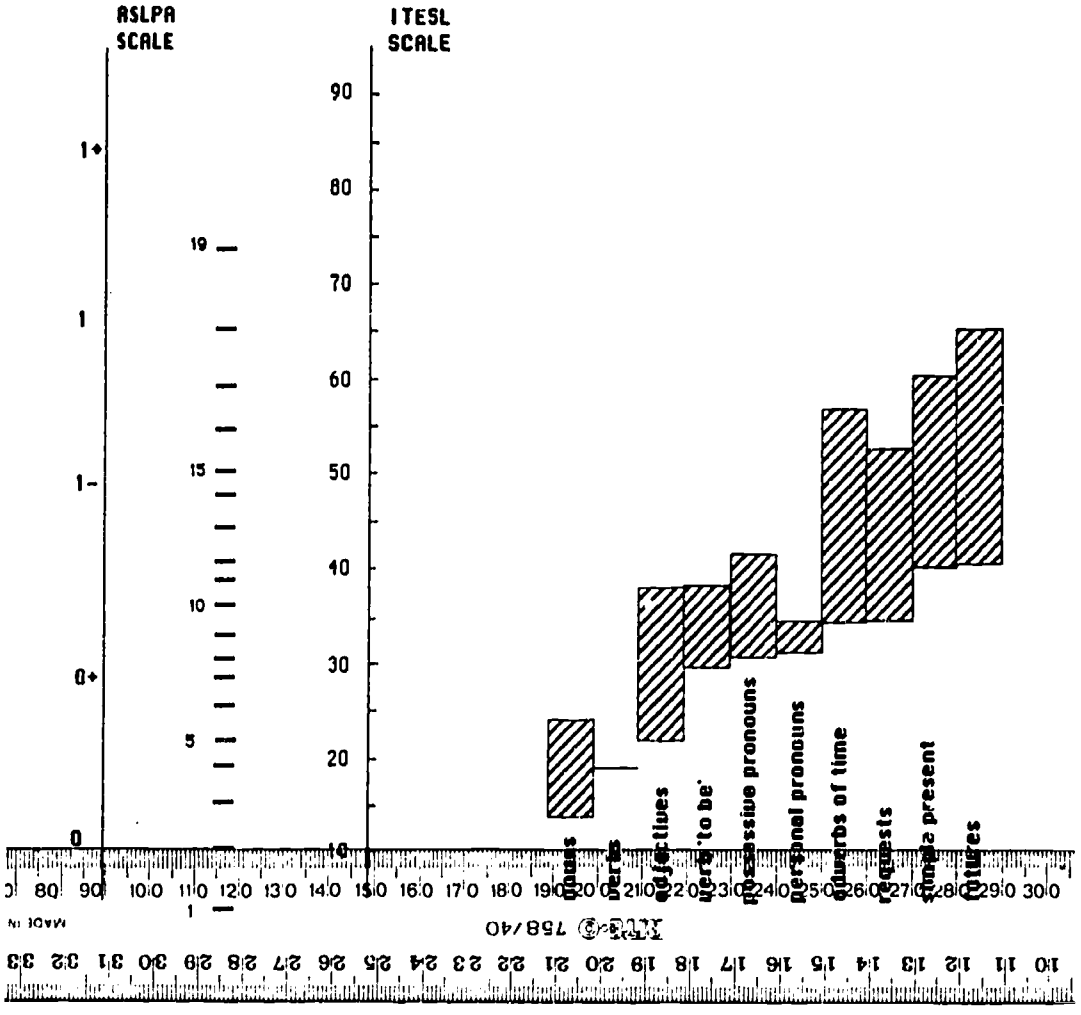


Figure 11 Identifying Regions of Most Probable Responses

All items except the "verbs" item have shaded regions above which the expected is two and below which the expected is zero. The verbs item does not have a shaded region because there is no score on the ITESL scale that leads to an expected score of one. Students who score below 17 on the ITESL have an expected score of zero. Students with a score above 17 would have an expected score of two. Some students may score one on the item but a score of one is never the most likely response. Further discussion on item plots can be found in Griffin, Adams, Martin and Tomlinson (1985).

Diagnosis

Plotting the regions of most probable response enables the ITESL score to be used for the diagnosis of specific strengths and weaknesses of individual students. This is one of the most important and powerful uses of the test.

To illustrate how the regions of most probable response can be used for diagnosis, two examples are shown below.

Example 1

First, consider a student who was administered Test 1 and had a raw score of 12 made up of scores to each item as shown in the first column of Table 6. A raw score of 12 gives the student an ITESL rating of 42 from which the student's most probable response to each item can be calculated (see Figure 12); these are reported in the second column of Table 6. The larger differences between the actual response score and most likely responses to the interview items can provide valuable diagnostic information.

This particular student appears to have scored below expectation on the first three questions but over scored on the possessive pronouns. The student's remaining responses are as expected. This indicates that the student may have difficulties in the area of basic vocabulary. This type of individual diagnosis should enable the teacher, or whoever has responsibility for placement, to provide more intensive individualised goals for the student.

The information available from Figure 12 and Table 6, indicates a partial mastery region at the end of the test. For "adverbs of time", "requests", "simple present" and "futures", the student has exhibited only part of an appropriate response; and these were the most likely types of response from that student. In general, the trend indicates that instructional time is required for the student to acquire and learn these structures in a variety of contexts.

Example 2

As a second example consider a student who was administered Test 3 and had a score of 13 made up of the responses in Table 7. A raw score of 13 corresponds to an ITESL of 62 (see Figure 13) for which the most probable responses are also recorded in Table 7. The discrepancies between the observed and most probable responses indicate that this student has problems with functional type questions but has a stronger mastery of the formal language structures. Details of the specific difficulties can be determined from the content of the items.

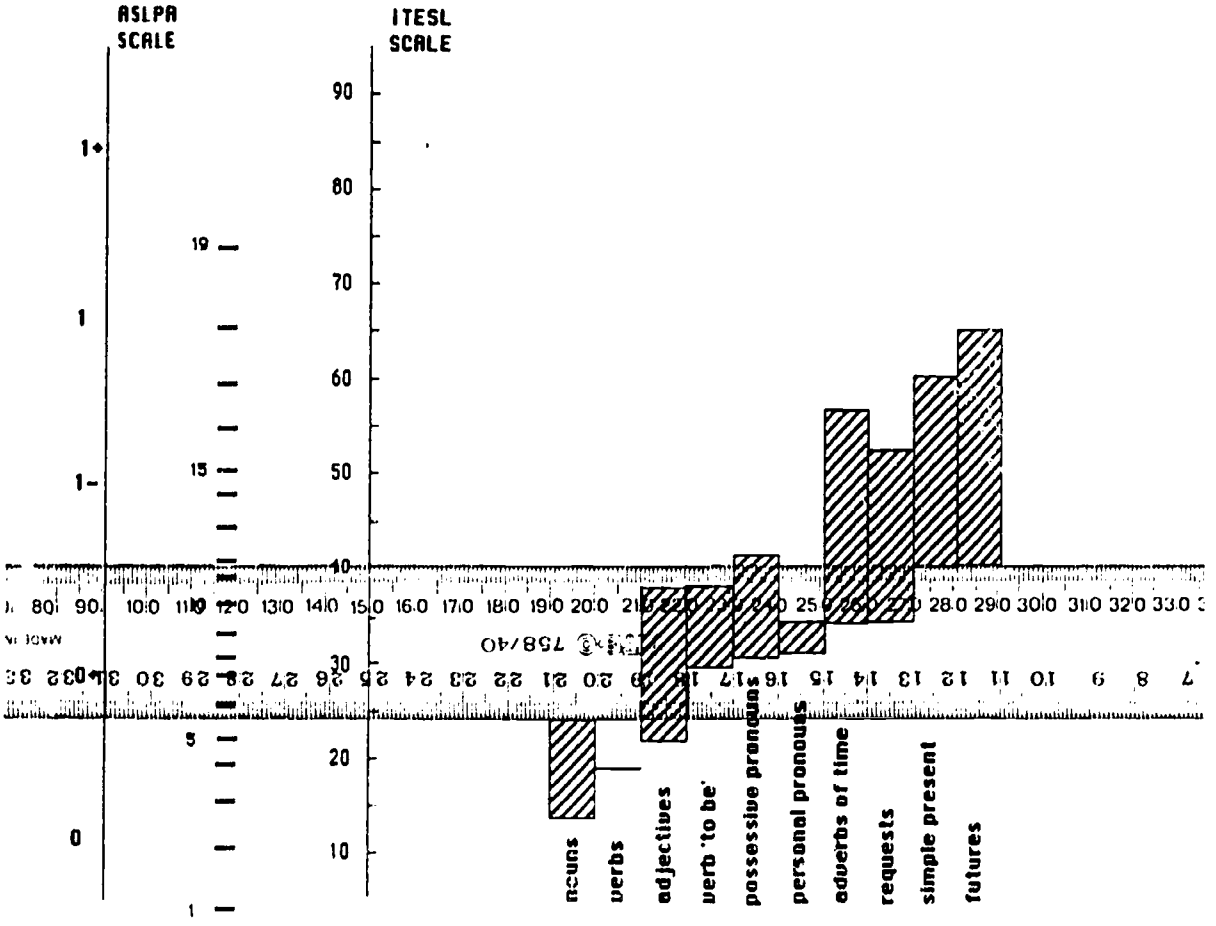


Figure 12 Most Probable Responses for Test 1

Table 6 Actual and Expected Responses for Test 1

Item	Actual score	Expected score
nouns	0	2
verbs	1	2
adjectives	1	2
verb 'to be'	2	2
possessive pronouns	2	1
personal pronouns	2	2
adverbs of time	1	1
requests	1	1
simple present	1	1
futures	1	1

It may be that the difficulties stem in this instance from a confusion of "left" versus "right", or from using simple prepositions in giving directions. The interviewer should note these specific difficulties in the space at the bottom of the score sheet. The second large difference between the actual score and most likely score is associated with "offers and invitations". This may arise from confusion about polite forms or requests. Again the interviewer should note these difficulties at the bottom of the answer sheet.

In general, examining the discrepancies between the actual and expected responses for a student can prove to be a very powerful diagnostic tool. When interpreting discrepancies between the actual and the expected responses, it is necessary to concentrate on the largest differences or on patterns. Small differences may occur due to rounding error and the use of a three-point scale only. Further, it should be noted that the totals of the actual and expected scores may not be equal. This is a result of rounding error when determining the expected scores.

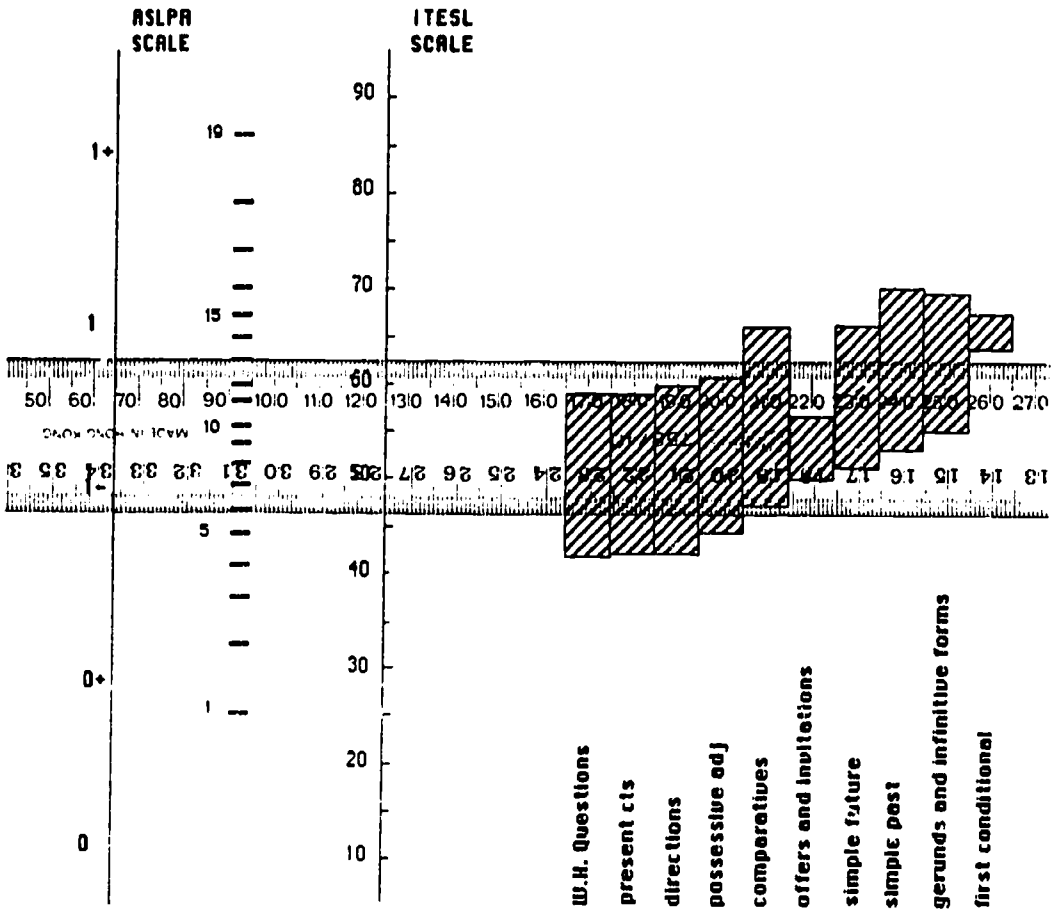


Figure 13 Most Probable Responses for Test 3

Conclusions

The possible uses of this project have been detailed in the pages of this report. In summary, these have been seen as the discrimination of oral proficiency through the application of technological advances made in the application of Rasch models in the area of language proficiency. In this study we have applied the Partial Credit Model, the most general and complex of the Rasch models. The application of this model is only now beginning to be investigated in a range of settings. This is believed to be its first application in the area of language development and, in particular to the speaking skill, which has perhaps been one of the most difficult of measurement areas.

Table 7 Actual and Expected Responses for Test 3

Item	Observed score	Expected score
W.H. Questions	1	2
present continuous	2	2
directions	0	2
possessive objectives	2	2
comparatives	2	1
offers and invitations	0	2
simple future	2	1
simple past	1	1
general and infinitive forms	2	1
first conditional	1	0

A range of uses have been indicated for the ITESL test that has been developed. These include: the detailed diagnosis of clients' specific strengths and weaknesses, the monitoring of the development of clients' oral proficiency and the placement of clients on the basis of oral proficiency.

On the basis of an examination of teachers' reports, the observation of teachers' practices and an examination of a large volume of literature, this study has taken a particular stance in the development of test objectives. While the adoption of this stance may be seen as controversial in some areas, the results of the test analysis have clearly supported the theoretical position adopted.

The testing procedures and technology that have been implemented in this study also have implications beyond this project. The methodology discussed and implemented could have important implications for a range of research in he language area. For example, the work of Dulay and Burt (1974a, 1974b) or Pienemann and Johnston (1985) could easily be validated with the application of a measurement approach similar to that adopted in this study. Furthermore, other dimensions in language proficiency that have been proposed may be tested and validated.

We hope that the use of the ITESL will assist the work of the AMEP in solving the problems which led to the generation of the research project.

PROPOSED RESEARCH PROJECT : EVALUATION IN THE AMEP (STAGE 2)

9605-05-05
TW DK
30 April 1984

GOAL: To undertake and report on a trial implementation of an evaluation model outlined in the paper "Evaluation in the Adult Migrant Education Program (Stage 1)" and develop mechanisms and testing instruments necessary for implementation across the AMEP.

OBJECTIVES

Over the term of the project the research officer will:

A. 1. Survey teacher practices in two specific venues to identify testing and assessment tools currently employed within AMES (Victoria).

2. Supplement the above base paper with a review of literature in the area of education program evaluation.

3. Review literature in the area of testing and assessment and identify a range of components which could contribute to the development of course specific tests as required by teachers.

4. Recommend, if available assessment tools are inadequate or inappropriate, the development of tools which meet the requirements of the program and describe these proposed tests in detail.

B. 1. Within a specific area of the AMEP, in close co-operation with centre staff, assist with the implementation of the model outlined in the above-mentioned paper. Implementation would involve:

a) developing curricula to meet the needs of specific groups expressed by way of behavioural objectives.

b) selecting students which fit the profiles used as a basis for the curricula developed.

c) assist staff in developing the expertise to judge whether or not students have met stated course objectives. This would involve staff developing course specific tests.

d) assist staff with modification of curricula, where appropriate.

2. Document a range of AMEP curricula for designated groups, define in terms of behavioural objectives.

3. Consider the application of computer technology for curriculum design, curriculum modification and course selection.

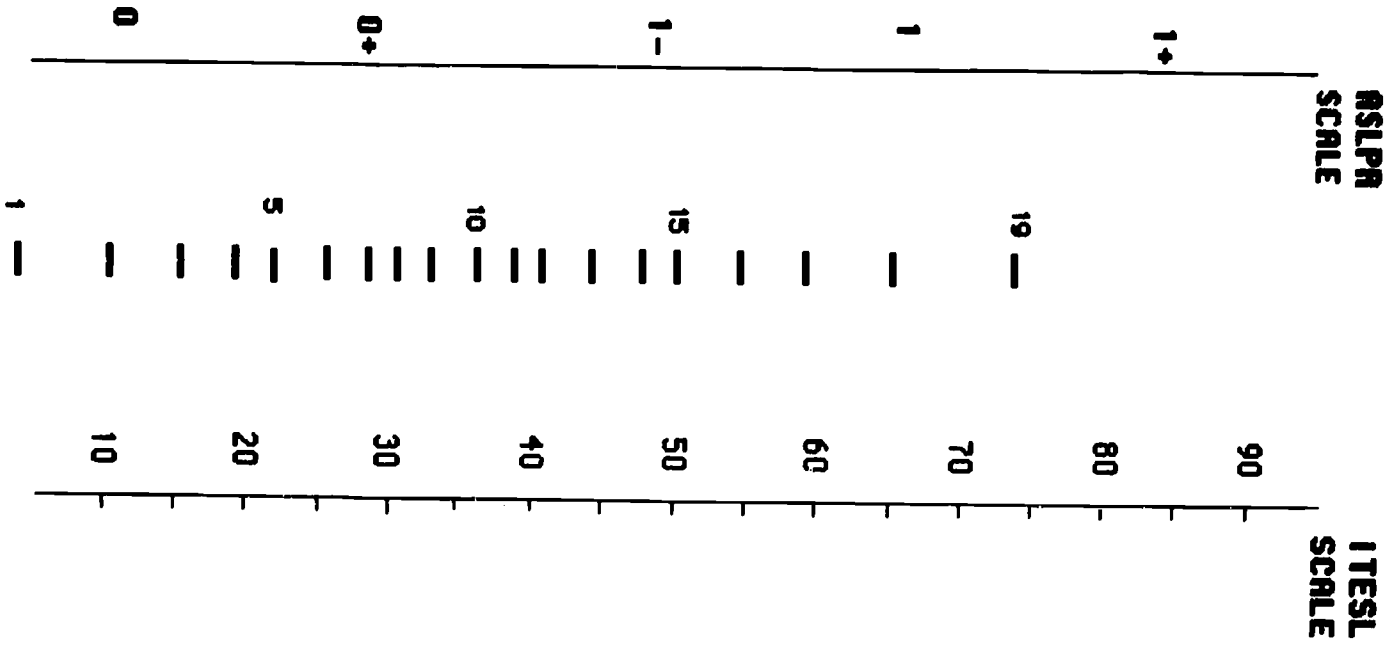
4. Advise on the practicality of a course approval committee to be responsible for approving new courses and for the modification of existing ones.

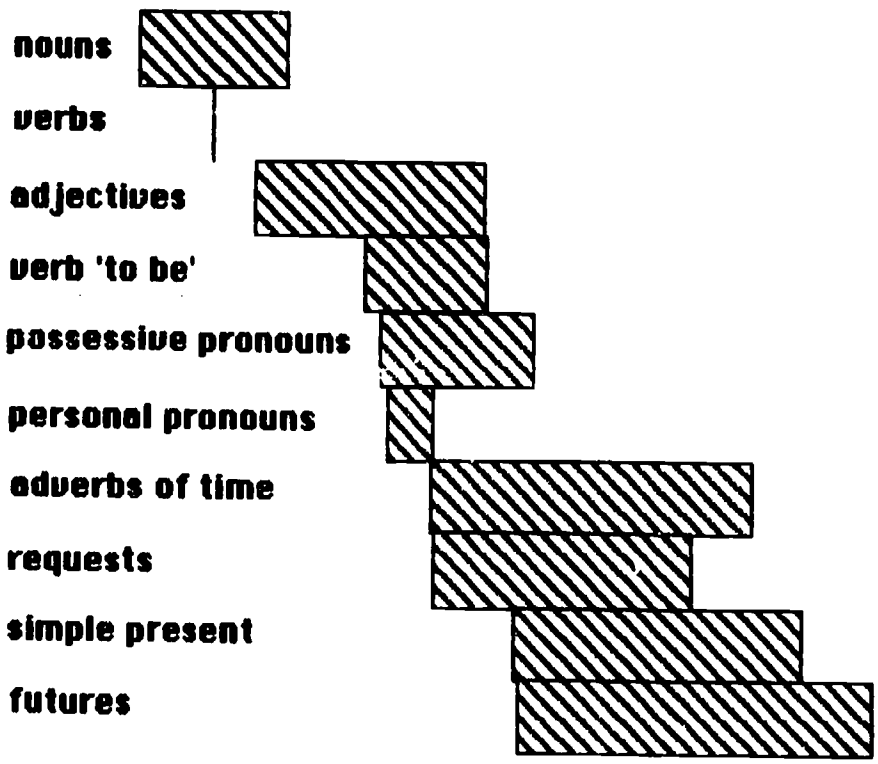
APPENDIX A

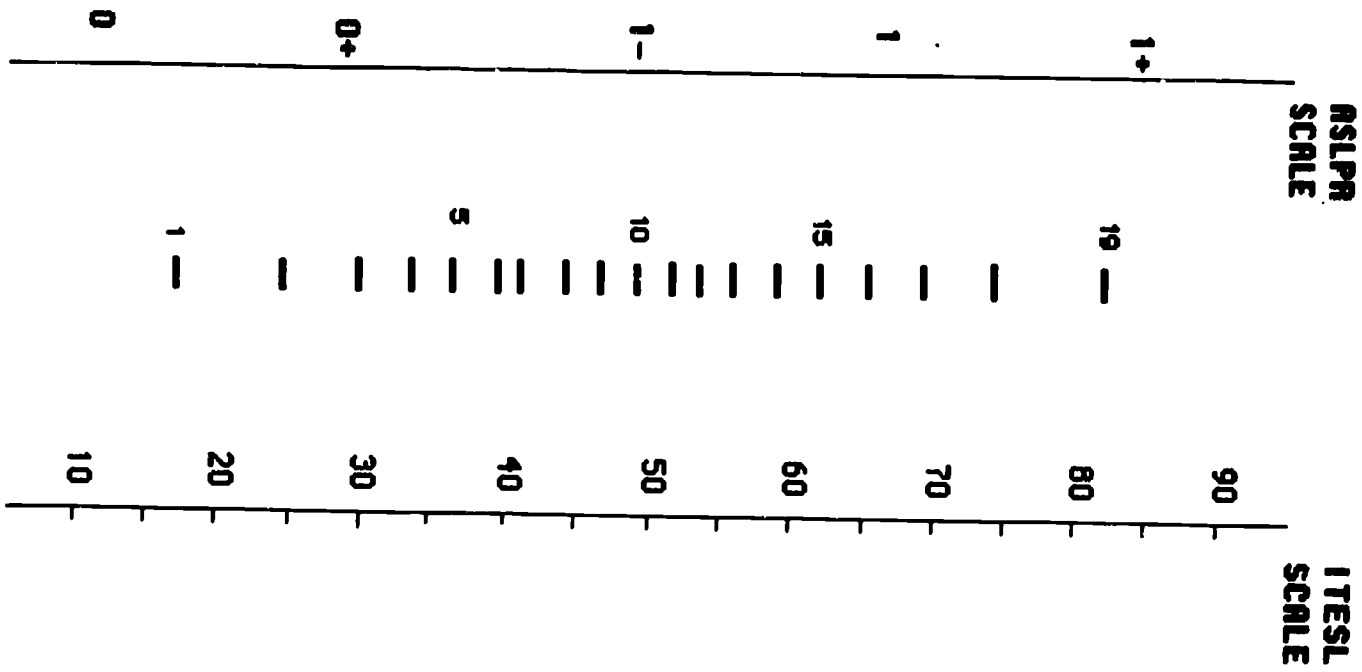
Prepared by Tim Walker
Executive Officer.

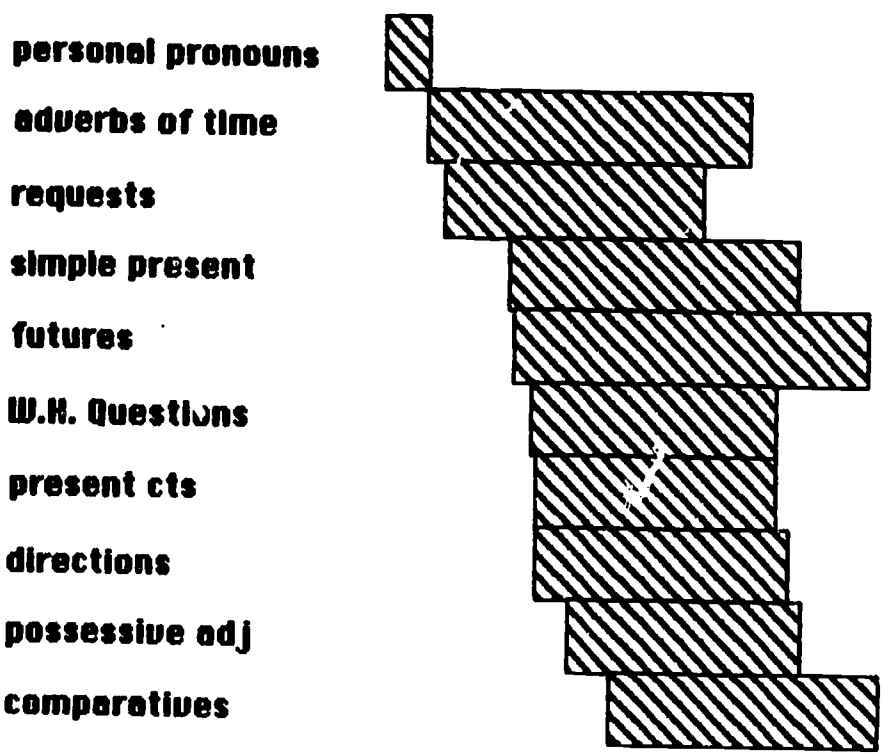
APPENDIX B

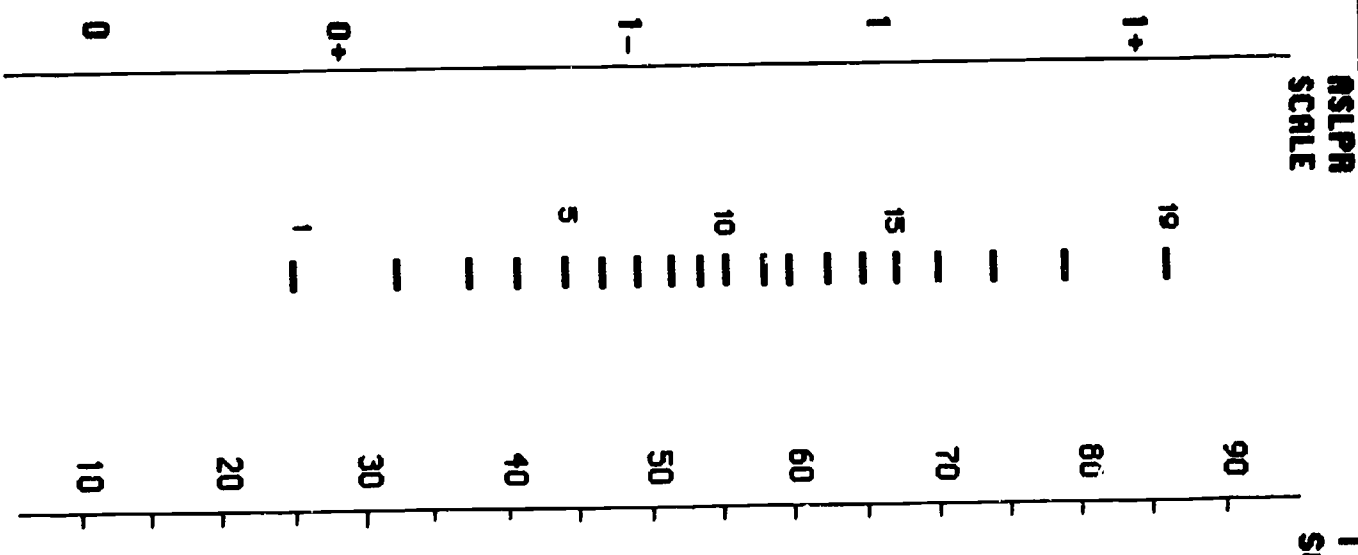
EQUIVALENCE TABLES

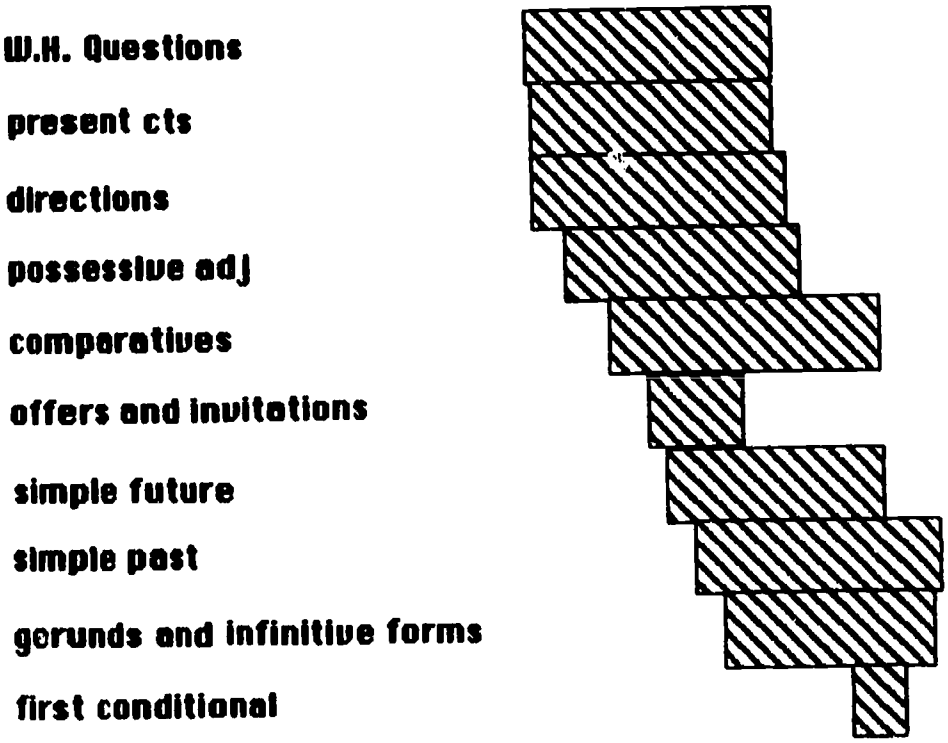





0
0+
1-
1
1+
ASLPR SCALE
1
5
10
15
19
10
20
30
40
50
60
70
80
90
ITESL SCALE


personal pronouns
adverbs of time
requests
simple present
futures
W.H. Questions
present cts
directions
possessive adj
comparatives


0
0+
1-
1
1+
ASLPR SCALE
1
5
10
15
19
10
20
30
40
50
60
70
80
90
TESL SCALE


W.H. Questions
present cts
directions
possessive adj
comperatives
offers and invitations
simple future
simple past
gerunds and infinitive forms
first conditional

REFERENCES

Adult Migrant Education Program of Victoria. 1983. Syllabus Guidelines. Melbourne: AMES.

Alderson, J.C. 1983. The cloze procedure and proficiency in English as a foreign language. In Oller, J.W. Issues in Language Testing Research. Rowley, Mass.: Newbury House.

Alkin, M.C. 1969. Evaluation Theory Development. Evaluation Comment. 2(1), 2-7.

AMES. 1984. "Evaluation in the Adult Migrant Education Program (Stage 4)". A discussion paper prepared by the Working Party. Victoria : AMES.

Andrich, D. 1975. The Rasch multiplicative binomial model: application to attitude data. Research Monograph No. 1. Measurement and Statistics Laboratory, Department of Education, University of Western Australia.

Andrich, D. 1978a. A binomial latent trait model for the study of Likert-scale attitude questionnaires. British Journal of Mathematical and Statistical Psychology. 31: 84-98

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

Andrich, D. 1978c. Application of a psychometric rating model to ordered categories which are scored with successive integers. Applied Psychological Measurements. 2: 581-594.

Andrich, D. 1980. Using latent trait measurement models to analyse attitudinal data: a synthesis of viewpoints. In Spearitt, D. Proceedings of the Invitational Conference on the Improvement of Measurement in Education and Psychology. Hawthorn: Australian Council for Educational Research.

Andrich, D. 1982. An extension of the Rasch model for ratings providing both location and dispersion parameters. Psychometrika. 47 : 105-113.

Australian Council for Educational Research. 1976. Tests of English for Migrant Students. Hawthorn: ACER.

Australian Council for Educational Research. 1981. ACER Listening Tests for 10-year-old and 14-year-olds. Melbourne: ACER.

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

Birnbaum, A. 1968. Some latent trait models and their use in inferring an examinee's ability. In, Lord, F. and Novick, M. Statistical Theories of Mental Test Scores. Reading, Mass.: Adison Wesley.

Bloom, B.J. et al. 1956. Taxonomy of Educational Objectives I. Cognitive Domain. New York: David McKay.

Bock, R.D. 1972. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika. 37: 29-65.

Brindley, G. 1984. Needs analysis and objective setting in the Adult Migrant Education Program. Sydney, New South Wales: Adult Migrant Education Services.

Brumfit, C. 1981. Teaching the General Student in Johnson, K. and Morrow, K. (eds) Harlow, Essex: Longman.

Brunton, C. and Gibbons, J. 1976. "Language Testing Project: Testing Oral Production", unpublished document. Department of Education, Lancaster University, available from J.P. Gibbons, School of Education, University of Hong Kong.

Burrill, L. 1976. The development of the standardised test. Measurement Newsletter No. 24. New York: The Psychological Corporation.

Canale, M. and Swain, M. 1980. Theoretical bases of communicative approaches to second language teaching and testing. Applied Linguistics. 11(1): 1-47.

Carroll, J.B. 1983. Psychometric theory and language testing. In Oller, J.W. Issues in Language Testing Research. Rowley, Mass.: Newbury House.

Choppin, B. 1982. The use of latent trait models in the measurement of Cognitive Abilities and Skills. In Spearitt, D. (ed) The improvement of Measurement in Education and Psychology. Melbourne: Australian Council for Educational Research.

Clark, J.L. 1979. The syllabus - what should the learner learn? Audio-visual Language Journal. 17(2)

Cronbach, L.J. 1963. Course improvement through evaluation. Teachers College Record. No.64.

Cronbach, L.J. 1970. Essentials of Psychological Testing. New York: McGraw Hill.

Davies, A. 1982. Language testing. In Kinsella, V. (ed) Surveys 1 and 2. Eight State-of-the-art Articles on the Key Areas in Language Teaching. Cambridge: Cambridge University Press.

Douglas, G.A. 1978. Conditional maximum - likelihood estimation for multiplicative binomial response model. British Journal of Mathematical and Statistical Psychology. 31: 73-83.

Douglas, G.A. 1982. Issues in the fit of data to psychometric models. Educational Research and Perspectives. 9,1: 32-43.

Dulay, H. and Burt, M. 1974a. Natural sequences in child second language acquisition. Language Learning. 24: 37-53.

Dulay, H. and Burt, M. 1974b. Errors and Strategies in child second language acquisition. TESOL Quarterly, 8: 129-136.

Dunn, L.M. and Markwardt, F.C. 1970. Peabody Individual Achievement Test (PIAT). Circle Pines, Minnesota: American Guidance Service Inc.

Duran, R.P. 1984. Some implications of communicative competence research for integrative proficiency testing. In Rivera, C. (ed) Communicative Competence Approaches to Language Proficiency Assessment: Research and Application. Clevedon: Multilingual Matters Ltd.

Eisner, E. 1975. Instructional and expressive objectives. In Golby et al. Curriculum Design. London: Croom Held Ltd.

90.

Farhady, H. 1979. The disjunctive fallacy between discrete-point and integrative tests. TESOL Quarterly. 13(3): 347-357.

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

Fortington and Cartwright. 1985. English as a Second Language. Revised Syllabus Guidelines. Victoria: Adult Migrant Education Services.

Galbally, F. (Chairman) 1978. Migrant Services and Programs, Report of Post Arrival Programs and Services for Migrants. Canberra: AGPS.

Goldstein, H. 1979. Consequences of using the Rasch model for educational assessment. British Education Research Journal. 5: 211-20.

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

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

Guttman, L. 1950. The basis for scalogram analysis. In Stoutter et al. Measurement and Prediction. New York: Wiley.

Hammond, R.L. Evaluation at the local level. Tucson, Arizona: EPIC. Evaluation Centers (undated mimeo).

Hastings, T. 1969. Keith and kin of educational measures. Journal of Educational Measurement. 6, 127-130.

Higgs, T.V. (ed) 1984a. Teaching for Proficiency, the Organising Principle. Lincolnwood, Illinois: National Textbook Company.

Higgs, T.V. 1984b. Language teaching and the quest for the Holy Grail. In Higgs, T. (ed) Teaching for Proficiency, the Organising Principle. Lincolnwood, Illinois: National Textbook Company.

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

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

Ilyin, D. 1976. Ilyin Oral Interview. Rowley, Mass.: Newbury House.

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

Izard, J. 1981. "The Robustness of the Rasch Model", unpublished paper. Melbourne: Australian Council for Educational Research.

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

Johnston, M. 1985a. Syntactic and Morphological Progressions in Learner English. Department of Immigration and Ethnic Affairs, 1985.

Johnston, M. 1985b. "Second Language Acquisition Research in the Adult Migrant Education Program". Sydney, New South Wales: Adult Migrant Education Services (mimeo).

Lado, F. 1961. Language Testing: the Construction and Use of Foreign Language Tests. London: Longman.

Liskin-Gasparro, J.E. 1984a. ETS Oral Proficiency Testing Manual. Princeton, N.J.: Educational Testing Service.

Liskin-Gasparro, J.E. 1984b. The ACTFL Proficiency Guidelines: Gateway to Testing and Curriculum. Foreign Language Annuals. 17(5).

Lumsden, J. 1976. Person Reliability. Applied Psychological Measurement. 4: 477-482.

Masters, G. 1980. A Rasch Model for Rating Scales. Doctoral Dissertation, University of Chicago.

Mager, R.F. 1973. Measuring Instructional Intent. California: Fearon Publishers.

Masters, G. 1982. A rasch Model for partial credit scoring. Psychometrika. 47: 149-174.

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

Masters, G. and Wright, B.D. 1984. The essential process in a family of measurement models. Psychometrika. 49: 529-544.

Masters, G., Wright, B. and Ludlow, L. 1980. CREDIT. A Computer Program for Analysis Data Scored in Ordered Response Categories. University of Chicago.

Macdonald, B. 1971. "The Evaluation of the Humanities Curriculum Project: a Wholistic Approach". Paper presented to the AERA Annual Meeting.

Nunan, D. 1985. "Recent Trends in Language Syllabus Design". Report prepared for the PDSC, Department of Immigration and Ethnic Affairs.

Oller, J.W. 1979. Language Tests at School, A Pragmatic Approach. London: Longman.

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

Oller, J.W. 1983b. Evidence for a general language proficiency factor: an expectancy grammar. In Oller, J.W. (ed) Issues in Language Testing Research. Rowley, Mass: Newbury House.

Page, B. 1979. Notions, functions and threshold levels: a review of the significance for language teachers of the work of the Council of Europe. Audio-Visual Language Journal. 17(2): 115-121.

Parlett, M. and Hamilton. 1976. Evaluation as Illumination. In Tawney, (ed) Curriculum Evaluation Today: Trends and Implications, Schools Council Research Studies. London: Macmillan.

Paulston, C. 1981. Notional syllabuses revisited: some comments. Applied Linguistics. 11(1): 93-95.

Pienemann, M. 1983. Learnability and syllabus construction. In Hyltestam, K. and Pienemann, M. (eds) Modelling and Assessing Second Language Development. Clevedon, Avon: Multilingual Matters. (to appear)

Pienemann, M. and Johnston, M. 1984. "Towards an explanatory model of language acquisition." Paper presented at the 9th Applied Linguistics Assocation of Australia Annual Congress, Alice Springs, 29 August - 2 September.

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

Provus, M. 1969. Evaluation of ongoing programs in the public school system. In Tyler, R. (ed). The 68th Yearbook of the NSSE, Part II, Educational Evaluation: New Roles, New Means. Chicago, Illinois: View Chicago Press.

Rasch, G. 1960 (1980) Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedegogiske Institut, 1960. (Chicago: University of Chicago Press)

Rivera, C. (ed) 1984. Communicative Competence Approaches to Language Proficiency Assessment: Research and Application. Clevedon: Multilingual Matters Ltd.

Samejima, F. 1969. Estimation of latent ability using response patterns of graded scores. Psychometrika, Monograph Supplement No. 17.

Spolsky, B. 1975. Concluding statement. In Jones, R.L. and Spolsky, B. (eds) Testing Language Proficiency. Arlington, Va.: Center for Applied Linguistics.

Spolsky, B. 1978. Linguistics and language testing. In Spolsky, B. (ed) Advances in Language Testing. Arlington, Va.: Center for Applied Linguistics.

Stake, R.E. 1967. The countenance of educational evaluation. Teachers College Record. 68: 523-540.

Stake, R.E. 1973. "Program evaluation, particularly responsive evaluation." Paper presented at New Trends in Evaluation, Gotesborg, Sweden, October.

Stufflebeam, D.L. Evaluation as enlightenment for decision making. An address delivered at the Working Conference on Assessment Theory. The Association for Supervision and Curriculum Development, Sarasota, Florida an. 1968.

Scriven, M. 1967. The Methodology of Evaluation. Perspectives of Curriculum Evaluation. AERA Monograph series on Curriculum evaluation, No.1, Chicago, Illinois: Rand McNally.

Tawney, D. (ed) 1976. Curriculum Evaluation Today. Macmillan.

Tosi, A. 1984. Immigration and Bilingual Education. Oxford: Pergamon Press.

Tyler, R.W. 1958. The evaluation of teaching. In Cooper, R.M. (ed) The Two Ends of the Log. Minneapolis, Minn.: University of Minnesota Press.

Vollmer, H.J. and Sang, F. 1982. Competing hypotheses about second language ability: a plea for caution. In Oller, J.W. Issues in Language Testing Research. Rowley, Mass.: Newbury House.

Wilkins. 1981. Notional syllabuses revisited. Applied Linguistics. 2(1): 83-89.

Wiseman, L. and Pidgeon, A. 1970. Curriculum Evaluation. The National Foundation for Educational Research.

Worthen, B. 1977. "Evaluation workshop materials used at the Australian Association for Research in Education Annual Conference". November 1977.

Wright, B.D. and Bell, S.R. 1984. Item banks: what, why, how. Journal of Educational Measurement. 21: 331-346.

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

Wright, B.D. and Stone, M.H. 1980. Best Test Design. Chicago: MESA Press.