

A PROPOSAL FOR CREATING TEST ITEM BANKS  
AT UNIVERSITAS TERBUKA

by

Sumedi P. NUGRAHA  
BA, SP, Gadjah Mada University

A PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF  
THE REQUIREMENTS FOR THE DEGREE OF  
MASTER OF EDUCATION  
in the Faculty  
of  
Education

© Sumedi P. NUGRAHA 1987

SIMON FRASER UNIVERSITY

October, 1987

All rights reserved. This work may not be  
reproduced in whole or part, by photocopy  
or other means, without permission of the author.

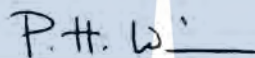
APPROVAL

Name : Sumedi Priyana Nugraha  
Degree : Master of Education  
Title of Project : A Proposal For Creating Test Item Banks At  
Universitas Terbuka

Examining Committee



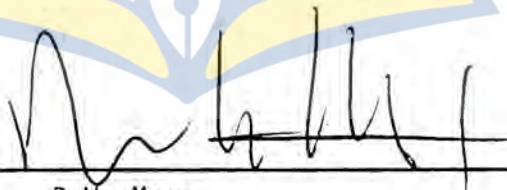
T. J. O'Shea  
Senior Supervisor



P. H. Winne  
Professor



N. Nasution  
Assistant Dean of Education  
Universitas Terbuka



R.W. Marx  
Professor

Date approved 9th December, 1987

### Abstract

Universitas Terbuka uses a multiple-choice format test as the only means for assessing students' achievement. Universitas Terbuka does not statistically analyze its item collections although item analysis is a crucial process in any objective test development. In this study, I propose the establishment of an item bank model at Universitas Terbuka. The main task in developing the item bank system is to carry out test item analyses. In this study I analyzed test items using two approaches: classical item analysis, and the Rasch Model.

The main features of classical item analysis include p-value, discriminating power, and distractor analysis. The purpose of doing this analysis was to delete poor items and items were retained if  $r_{pbis} > 0.2$  and all options were chosen by some students. Furthermore, only one distractor was allowed to have a positive  $r_{pbis}$  up to 0.05. Classical item analysis showed only 38% of items on the December 1986 and 54% of items on the May 1987 English examination were satisfactory. Furthermore, 33% of items in December 1986 and 70% of items in the May 1987 Mathematics

examination were satisfactory. The rest of the items were not suitable for assessing student achievement because they had poor characteristics.

The Rasch Model provides a different approach to item analysis. This model is very useful for adjusting all items in the collection onto a common difficulty scale, especially when teachers want to create different tests to measure the same objectives. In this study a number of items common to the December 1986 and May 1987 examination were used to adjust all the remaining items onto a common difficulty scale.

To benefit from the calibrated items, all items need to be entered into an item bank in order that item collections are easy to be retrieved. In the final chapter of this study, I propose a systematic procedure for creating, field-testing, analyzing, calibrating items, and constructing an item bank to manage test production.

### Acknowledgment

This project was completed with the assistance of members of the Faculty of Education at Simon Fraser University in Canada and at Universitas Terbuka in Indonesia. I wish to thank to all of those who provided me with help. I wish also to thank:

Dr. Thomas O'Shea, as the first committee member, for helping me when I had some difficulties, and for supporting me (at a distance).

Dr. Phil Winne, as the second committee member, for contributing excellent ideas.

Mr. Nuhi Nasution, MA, as the third committee member, for helping me to do item analyses at Universitas Terbuka.

John Anderson, Ph.D, as a consultant for evaluation at Universitas Terbuka, for reviewing my project during his visit to Indonesia.

I hope this project will be of use to people who work in developing test items, and especially for those interested in developing item banks at Universitas Terbuka.

Table of Contents

	page
Title Page .....	i
Approval Page .....	ii
Abstract .....	iii
Acknowledgment .....	v
Table of Contents .....	vi
List of Tables .....	ix
List of Figures .....	x
 CHAPTER I: INTRODUCTION	
Universitas Terbuka .....	1
History .....	1
Organization .....	3
Academic Programs .....	4
Students .....	6
Faculty of Teacher Education and Pedagogy .....	7
The Examination Process Centre .....	8
Item Bank Development .....	10
Problem Identification .....	11
The Goals of the Project .....	13
Significance of the Study .....	14
Definitions of Terms .....	15
The Organization of the Project .....	17
 CHAPTER II: REVIEW OF LITERATURE	
Distance Education .....	19
Background of Distance Education .....	19
Characteristics of Distance Education ...	21
Student Evaluation .....	22

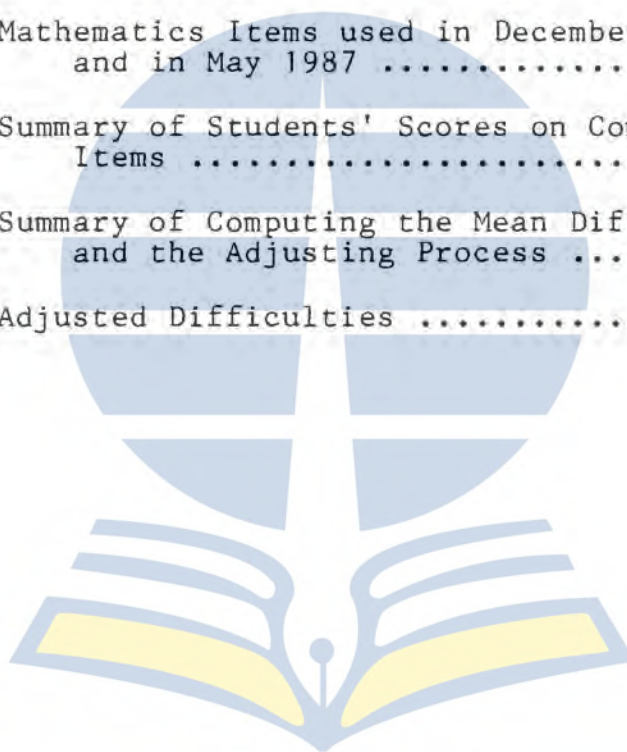
Item Analyses .....	24
Classical Item Analysis .....	24
Item Difficulty .....	24
Item Discrimination .....	26
Distractor Analysis .....	29
Rasch Model for Item Analysis .....	30
Estimating Item Difficulty and Person Ability .....	36
Item Bank Theory.....	37
Definition of Item Banks.....	37
Item Bank Types .....	40
Item Bank Functions.....	41
Test Development .....	41
Test Production .....	42
Item Maintenance .....	42
 CHAPTER III: METHODOLOGY	
ITEMAN and RASCAL .....	44
Preparing Data Files .....	45
Interpreting the Analysis .....	46
ITEMAN .....	46
RASCAL .....	48
Item Analysis Report .....	50
Collecting Student Scores .....	51
Analyzing Items .....	52
Classical Item Analysis .....	52
Rasch Model Item Analysis .....	56
Adjusting Item Difficulties to a Common Scale .....	60
 CHAPTER IV: PROPOSED ITEM BANK SYSTEM	
Items Development .....	66
Item Format .....	69
The Use of the Item Bank .....	72
Item Maintenance .....	73
Test Assembling .....	74
Test Administration, Scoring, and Reporting ..	75

System .....	76
CHAPTER V: DISCUSSION AND RECOMMENDATIONS	
Discussion .....	79
Item Analysis .....	79
Item Bank .....	81
Recommendations .....	82
Item Analysis .....	83
APPENDICES	
Appendix A: The Structural Organization of Universitas Terbuka .....	85
Appendix B: Sample of Raw Data of Student Responses in the English Examination held in May 1987 ....	87
Appendix C: Sample of the Results of Item Analyses using the Classical Model .....	89
Appendix D: Results of Item Analysis of the Reduced Item Collection using the Classical Model .....	91
Appendix E: Results of Item Analyses using the Rasch Model .....	119
Appendix F: Questions to be answered in designing Item Banking Systems ..	125
References .....	134



List of Tables

	page
Table 1-1 The Distribution of Universitas Terbuka's Students .....	7
Table 2-1 Evaluation of Discrimination Index .....	29
Table 3-1 Summary of Items Having Borderline Distractors .....	54
Table 3-2 English Items used in December 1986 and in May 1987 .....	58
Table 3-3 Mathematics Items used in December 1986 and in May 1987 .....	58
Table 3-4 Summary of Students' Scores on Common Items .....	59
Table 3-5 Summary of Computing the Mean Difference and the Adjusting Process .....	61
Table 3-6 Adjusted Difficulties .....	62



List of Figures

		page
Figure 1-1	The Structure of Universitas Terbuka Courses .....	6
Figure 2-1	The Rasch Model Characteristic Curve .	36
Figure 3-1	An Example of Data Layout .....	46
Figure 3-2	The Item Statistics and the Alternative Statistics .....	47
Figure 3-3	ITEMAN Summary Statistics .....	48
Figure 3-4	RASCAL Output when the Difficulty Scale is selected .....	49
Figure 3-5	The Selected Items Based on Classical Item Analysis .....	56
Figure 4-1	Example of an Examination Matrix .....	68
Figure 4-2	An Item Bank System for Universitas Terbuka .....	77



## CHAPTER I

## INTRODUCTION

Universitas TerbukaHistory

Universitas Terbuka is a new distance university situated in Indonesia, established in 1984. The main purpose of this institution is to accommodate more students wishing to attend university. The other purposes are, first, to provide enhancement for secondary school graduates, both for those who have work and those who do not. Secondly, the teaching/ learning process will improve through the development of learning materials, and the mastery and the application of educational technology will be augmented (Ministry of Education and Culture, 1984).

The idea of developing an Open University in Indonesia, which will be referred to as Universitas Terbuka, began ten years ago when the Indonesian Government implemented the Republic of Indonesia's Second Five-Year Plan known as Repelita II, 1974-78 (Depdikbud, 1984). The President of the Republic of Indonesia decided to establish Universitas Terbuka, with the Presidential Decree No. 41, on June 11, 1984.

Universitas Terbuka was also founded on the basis of Government Regulation No. 5, 1984. The organization of Universitas Terbuka was determined by Decree of the Minister of Education and Culture No. 0389/O/1984 dated 27 August 1984 (Ministry of Education and Culture, 1984), and Universitas Terbuka was stipulated as the 45th public university in Indonesia.

Universitas Terbuka has one central office, which is located in Jakarta, 32 Distance Learning Program Units ("Unit Program Belajar Jarak Jauh", or UPBJJ for short), and a number of learning centers spread out in Indonesia. The central office is primarily responsible for academic administration, and for the development, provision and distribution of learning materials, and learning media and student evaluation.

The Distance Learning Program Units carry out the management of the teaching-learning process in each region where Distance Learning Program Units are located. More specifically, the tasks of the Distance Learning Program Units are to serve students, to help the central office in managing general administration, and to assist with examination activities. In doing these activities, the local public university provides

some facilities.

The learning center is intended for professional education programs. For instance, it provides media to guide students in the teaching-learning process, and it holds tutorials and provides a place for conducting experiments.

The funds for operating Universitas Terbuka are obtained from the Government through the National Budget as well as from the students, through their tuition fees. Funds also come from private sources and foreign aid.

#### Organization

Based on the Presidential Decree No. 41, 1984, Universitas Terbuka has the following hierarchy: The Rector is the leader who is supported by three assistants: the first assistant (Purek I) deals with academic matters, research, and services to the community; the second (Purek II) handles the general university administration; and the third (Purek III) is responsible for student affairs.

To maintain organizational activities, Universitas Terbuka has two bureaus: the general administration bureau and the academic and student affairs

administration bureau. There are also four centres which support Universitas Terbuka activities: (1) the production centre for educational media, information, and data processing, (2) the examination process centre, (3) the Distance Learning Program Unit, and (4) The Centre for Research and Service to the Community.

Four faculties which are elements of Universitas Terbuka are:

1. The Faculty of Teacher Education and Pedagogy
2. The Faculty of Economics
3. The Faculty of Social and Political Science
4. The Faculty of Mathematics and Natural Sciences

Besides these structural units, there are also non-structural units which function as complementary elements to the university, for example, the University Senate and The Advisory Council. Appendix A depicts the structure of the organization of Universitas Terbuka.

#### Academic Programs

Study programs which are available at Universitas Terbuka are Strata I (S1), Diploma II (D2), and Diploma III (D3). However, not all faculties have Diploma programs. The S1 program (Sarjana Program) consists of

two options: the main program and the regular program. Students who attend the main program should have a minimum achievement index, or IP (indeks prestasi) in Indonesian, which is equal to 3.0. The achievement index is a combination of examination results and course credits. The achievement index ranges from 4, the highest, to 1, the lowest. The main program requires the students to complete courses which are equal to 160 credits. Besides that, these students have to write their own thesis. The thesis itself has 6 credits. On the other hand, students who enrol in the regular program should have a minimum achievement index which is equal to 2.0. The regular program requires the students to attain between 144 to 160 credits. There is no thesis, however.

Figure 1-1 describes the process of determining which program a student enters. It also delineates the anticipated program for students who plan to continue their study into the Master Degree (Strata II, or S2 for short) if they are willing to take more courses which are equal to 165 credits.

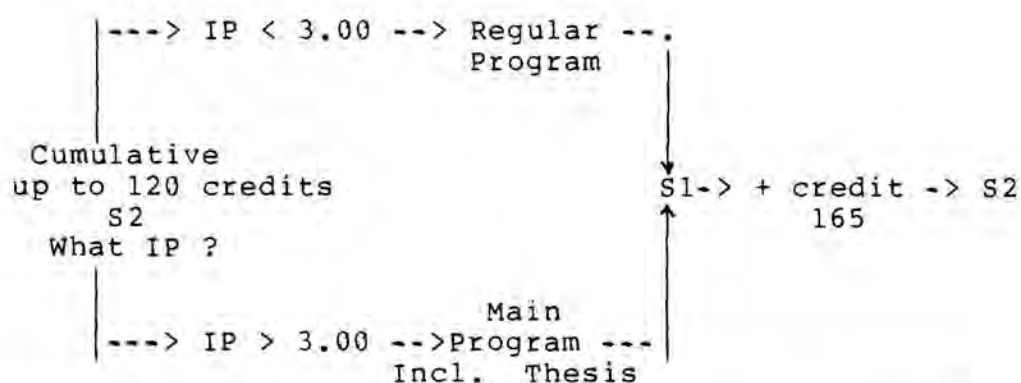


Figure 1-1 The Structure of Universitas Terbuka Courses

The D2 and D3 programs require between 80 and 90 credits and between 110 and 120 credits, respectively. These programs require that the students attain the minimum cumulative achievement index of 1.75. The functions of the Diploma program are to increase teaching competence.

### Students

As of April 1987 there were 64 342 students registered at Universitas Terbuka. They were enrolled in four faculties in the S1, D2, or D3 program as shown in Table 1-1.

The Faculty of Education differs from the other faculties in that it does not enrol students who have graduated directly from Senior High School level (SMA).



Table 1-1  
The Distribution of Universitas Terbuka's Students

Academic Program	Study Program	Number of Students
1. Indonesian Language	D2	949
2. English Language	D2	429
3. Natural Science	D2	712
4. Social Science	D2	814
5. Non-Formal Education	D2	48
6. Mathematics	D2	1 031
7. Pancasila Moral Education	D2	322
8. Sports and Health	D2	83
9. Indonesian Language	S1	717
10. Biology	S1	889
11. Chemistry	S1	601
12. Tax	D3	544
13. Public Administration	S1	25 906
14. Business Administration	S1	8 429
15. Development Administration	S1	1 003
16. Economics and Development Study	S1	11 181
17. Management	S1	4 140
18. Mathematics	S1	794
19. Applied Statistics	S1	3 560
20. English Language	S1	1 194
21. Physics	S1	472
22. Mathematics	S1	524
Total		64 342

Rather, it takes in teachers who have been working at Junior High School (SMP) or SMA, and who have completed their SMA.

#### Faculty of Teacher Education and Pedagogy

There are six academic programs for S1 available at the Faculty of Teacher Education and Pedagogy: (1)

English, (2) Chemistry, (3) Mathematics, (4) Biology, (5) Indonesian Language, and (6) Physics. The Diploma II programs are (1) Non-Formal Education, (2) Natural Sciences, (3) Indonesia Languages, (4) Mathematics, (5) Pancasila Moral Education, (6) English Language, (7) Social Science, and (8) Sports and Health.

#### The Examination Process Center

The Examination Process Center is one of the technical implementation units at Universitas Terbuka. This unit works with the administration of assessment for all faculties at Universitas Terbuka, and includes four main activities. First, it prepares examination materials for all courses at Universitas Terbuka. This activity includes collecting test items from four faculties, copying them onto 5.25 inch floppy disks, and printing them out as test booklets. The second activity is to coordinate examination activities. To assist in administering examinations at each Distance Learning Program Unit, the Examination Process Center publishes manuals for test administration. The third activity is to grade the students' achievement. In grading the students' achievement, the Examination Process Center works with the Computing Unit to scan

the answer sheets and grade their achievement. Finally, it statistically analyzes test items after administering the tests. This last activity, however, has been done only twice, in January 1985 and May 1987.

Student evaluation is carried out in two forms: a take-home assignment and a final examination. The take-home assignment is completed during the learning process at each student's home. The answer sheet for doing the take-home assignment is sent out together with the modules. This take-home assignment is an optional part of the assessment of students' achievement scores. If the students want to send back the answer sheet to Universitas Terbuka, then both the take-home assignment and the final examination contribute to the student achievement scores with comparative values as follows: 20 percent for the take home assignment and 80 percent for the final examination. If they do not submit the take-home assignment answer sheets, then their achievement score is based only upon their final examination.

There are two examination periods each year, in March and October. The examinations are held on Sunday of the second and third weeks in the month of

examination. Examination times are from 7:30 AM to 4:45 PM at all the learning centers.

#### Item Bank Development

The item bank development at Universitas Terbuka is intended to supply items for composing tests (Depdikbud, 1984). The purposes of tests and examinations are as follows:

- . As a tool for measuring student performance.
- . As a feedback device to students as part of the teaching-learning process in the form of self-tests or feed back questions that are an integral part of the modules.
- . As one of the tools for evaluating the quality of instructional packages for tutorials (p.80).

In reality, the latter two programs are not now important, and the purpose of building an item bank at Universitas Terbuka is to develop items and to assemble tests to measure students' competency after the students learn the course material at a distance.

For the item bank each Faculty develops and maintains its own item collection. The Faculty uses file cards to store the items. A computerized item bank was begun for the examination in December 1986, and all items which had been used in those examinations

were stored as an item pool using the dbase III package program for the IBM PC.

#### Problem Identification

The critical problems in item bank management at Universitas Terbuka relate to item analysis and the creation of item sets to compose tests. The problems of item analysis consist of computing the validity and the reliability of each item, and also computing item difficulty, item discrimination and distractors analysis. Universitas Terbuka has not analyzed all of its item collections yet. What Universitas Terbuka has done is simply to create items, assemble and administer tests, and score and report the results (Anderson, 1985). Hambleton and Swaminathan (1985) say that the important stages in developing tests, especially in using item response theory, are field testing and selection of test items, and reliability studies. A good procedure for developing an item bank system is for the item bank developers to be primarily concerned with the item analysis process.

The second problem, which is part of the process of composing tests, deals with determining the difficulty level of each test. This is the most crucial stage in

setting tests. This problem is very similar to the first problem, in as much as it deals with item calibration. Item calibration is very important when assembling tests from an item bank. Using item response theory for item selection, several different tests of the same difficulty level may be constructed. This is very useful, as the evaluators can maintain the security of items and they can also make as many test variations as they want. From the students' point of view, they might not complain about unfairness in testing.

The Minister of Research and Technology of Indonesia has criticized the multiple-choice format which is used for student evaluation in most schools in Indonesia. He says that the multiple-choice test format is very bad for detecting academically qualified Indonesian people (Kedaulatan Rakyat, Sept, 86 p. 1, cols. 5-6). This statement poses a challenge for Universitas Terbuka which also uses the multiple-choice test format. The Minister has a valid objection if the process of developing items is done without using an adequate item analysis and without reviewing items periodically.

In this project, I analyze examinations in English and Mathematics courses of the Faculty of Education. More specifically, I am concerned with the Structure II (English Course) and the Statistics (Mathematics Course) of the D2 program. My concern with the English course relates to the problems involved with English as a Second Language (ESL) in Indonesia. Students have learned English since they were at Junior High School. However, the quality of their English at the University level is not satisfactory.

Another problem is with the Statistics. The students have weaknesses in understanding data which is presented using statistical symbols. To overcome these problems, the process of education should be improved. Improving the measurement tools in the assessment of English and Mathematics courses could help to promote the success in this learning process.

#### The Goals of the Project

This project has three goals. First, I will analyze the items which have been used for examinations in December 1986 and in May 1987 in the Structure II and Statistics courses for the Diploma II program. To analyze items, classical item analysis and the one-

parameter latent trait model or Rasch Model have been chosen. The software programs used to analyze the items are ITEMAN for analyzing items using the classical model, and RASCAL for analyzing items using the Rasch Model (Assessment System Corporation, 1986).

Second, I will develop an item bank model and procedure using classical item analysis for detecting poor items and the Rasch Model for calibrating items onto a common difficulty scale. Finally, I will suggest a procedure for assembling tests at Universitas Terbuka. These tests are based on the item bank which has been developed.

#### Significance of the Study

The assessment of student learning is an important component of education. Tests must be statistically reliable as well as being content valid. The procedures developed in this project will ensure that tests developed by Universitas Terbuka are both valid and reliable.

Classical item analysis and the Rasch Model will be used as models for the item calibration process. Classical item analysis is very useful in giving information about item difficulty, item discrimination,



and for doing analysis of item distractors.

Comparing classical test theory and the Rasch Model, the classical theory does not have the facility to compare between two subtests which measure the same thing. What makes the Rasch Model interesting is the invariability of its item parameter estimates. This means that the results of the item analyses using the Rasch Model can be adjusted for the characteristics of the tryout sample. Thus, there is a greater possibility of developing an item bank which can be used for larger groups of students without being affected by variations in the tryout group. Moreover, Wright and Stone (1979) state that the Rasch Model is very useful in managing the item collections in the bank when they are increasing, because the Rasch Model provides an excellent way to select items easily and to build more varied tests.

#### Definition of Terms

Distance education in this study refers to:

- separation of teacher and student
- influence of an educational organization, especially in the planning and preparation of learning materials

- use of technical media
- provision of two way communication
- possibility of occasional seminars
- participation in the most industrialised form of education (Keegan, 1983, p. 15).

The item banking system in this study is defined as a storage and retrieval system which consists of a set of items and a mechanisms to select items for the creation of tests (Anderson, 1985).

Item response model is defined by Hambleton (1979) as follows:

It is a model that supposes examinee performance on a test can be predicted (or explained) in term of one or more characteristics referred as to traits (p.14).

The term ability is also comprehensively described by Hambleton and Swaminathan (1985):

- Ability, e.g., "numerical ability," is the label that is used to describe what it is that the set of test items measures.
- The ability or trait can be a broadly defined aptitude or achievement variable (e.g., reading comprehension), a narrowly defined achievement variable (e.g., ability to multiply whole

numbers), or a personality variable (e.g., self concept or motivation).

- Construct validation studies are required to validate the desired interpretations of the ability scores.
- There is no reason to think of the "ability" as innate. Ability scores can change over time and they can often be changed through instruction (p.55).

So, principally what is meant by the label of ability is any dimension that is intended for measurement by testing. For instance, with a test which intends to measure a student's achievement, the ability is measured as an achievement score.

Item difficulty is defined as "the point on the ability scale at which an examinee has a 50 percent chance of answering the item correctly" (Renz & Rentz, 1978, p. 2).

Calibration refers to the process of estimating and evaluating two sets of parameters, and fitting them with the model (Renz & Rentz, 1978).

#### The Organization of the Project

This project is organized into five chapters. Chapter One presents an introduction to the topic, which consists of a description of Universitas Terbuka,

the identification of the research problem, the goals of the project, the significance of the study, and definition of terms.

Chapter Two contains a literature review. Three topics will be discussed: distance education, item analyses, and item banking.

Chapter Three reports the item analyses process. It also contains a description of item analyses using classical and Rasch Model item analyses.

Chapter Four suggests a model for an item banking system for use at Universitas Terbuka. It discusses the description of the proposal item bank.

Chapter Five gives some discussions and recommendations of the project.



## CHAPTER II

### REVIEW OF THE LITERATURE

This chapter, first of all, reviews distance education activities relevant to the study of the item bank at Universitas Terbuka. Secondly, it addresses item analyses. Two approaches of item analyses are introduced: the classical item analysis and the Rasch Model. Finally, it explores item bank systems which have been developed in the area of education.

#### Distance Education

##### Background of Distance Education

In discussing the background of distance education, I refer to the work of Holmberg (1982). He suggests that distance education exists because of a geographical distance between students and university in many countries. The other reason for developing a distance education institution is that there is an opportunity for people to pursue extra education while they work. There is also the need for individual kinds of study because of individual differences, different knowledge, and different study conditions (Holmberg, 1982). Distance education is supported by rapid developments in communication technology, such as

electronic conference and mail, and communication media, such as telephone, radio, and TV. These media supplement printed materials, which are the most evident distance education activities. Sewart, Keegan and Holmberg (1983) refer to this study activity as independent study, because the study is supported by printed material and many kinds of media.

Perraton (1983) also suggests that several assumptions underlie the development of distance education. He hypothesizes first that any form of media can be used to teach. Second, student attendance can be increased without adding teaching staff members. Third, distance education is cheaper than conventional education. Fourth, the economics achievable by distance education are a function of the level of education, size of audience, choice of media and sophistication of production. Finally, distance teaching can reach audiences who would not be reached by conventional means, such as face-to-face interaction between students and teachers.

Distance education institutions exist in both developing countries and developed countries. In developing countries, examples of distance education

institutions are The Allabama Iqbal Open University in Islamabad, Pakistan, and the Shukothai Thammathirat in Thailand. In developed countries, there are the British Open University in England, the Deakin Open University in Australia, and the Open Learning Institute in Canada. Seward (Seward et al., 1983) states that distance education is becoming more and more popular both in developing countries and developed countries. Moreover, according to Seward the reason for the popularity of distance education is that qualified teachers are very limited and competent teachers are very rare. As a result, distance education could be an alternative to traditional form of education.

#### Characteristics of Distance Education

Because learning activities in distance education are different from conventional schools, distance education has some characteristics which are pointed out by Seward, Keegan and Holmberg (1983):

- the development of self instructional study material, i.e. courses printed and/or recorded which may either be self-contained or of a study guided type relying on set texts.
- teaching at a distance by comments in writing, on the telephone or on audio cassettes on

students' work submitted.

- counselling and general support of students' work by the same distance-study media (pp.1-2).

These characteristics indicate that students learn the course materials by themselves. Coffey (see Woodley, 1979) states that the teaching orientation in open university is student centered. Students learn what they want, when they want, and how they want. However, it is also possible to make contact with the supervisors, tutors and counsellors both through face to face and distance communication.

#### Student Evaluation

Compared to the conventional university, according to Woodley (1979), the open university differs in administrative, educational, and informational procedures. In the administrative procedure, open university's students have a lot of freedom in determining their study program, study time and place to study. In educational procedures, they have opportunities to determine their learning sequences, methods and objectives. In informational procedures, they are provided with adequate publicity for the courses.



These differences between conventional and open universities result in differences in evaluating students' achievement, because face-to-face communication between teaching staff and students occurs in a very limited period of time. Most study time is through a distance. Thus, the only way to evaluate student achievement is by conducting written examinations. Holmberg (1982) argues that in developing assessment devices, an open university should avoid long essays, because the course developers should not waste either the student's or the tutor's time. According to Holmberg (1986), objective tests are good for assessing distance education institution students' achievement.

Tests need to be easily and quickly provided, and they should be also accurate in predicting student abilities. An item bank system could solve this requirement, because the item bank system is equipped with a mechanism for storage and retrieval. Items stored into the bank are calibrated, and for large item collections, teachers can assemble varied tests with minimum error in estimating student achievement. A key requirement in making an item bank work is that all

items are statistically analyzed and calibrated.

### Item Analyses

Good tests must contain good items. One way to demonstrate the quality of the items is by a statistical item analysis. Two approaches to statistical item analyses which can be used are classical item analysis, and the Rasch Model item analysis.

#### Classical Item Analysis

The important concepts of the classical item analysis are item difficulty, item discrimination (Kaplan & Saccuzzo, 1982), and item distractors (Popham, 1981). These characteristics determine whether test items are classified as good or poor.

#### Item Difficulty

The difficulty of an item, in classical item analysis, is defined as the proportion of students who answer the item correctly. For example, if 54% of students who take a Mathematics exam got the item correct, then, the difficulty level is 0.54. Thus the higher the difficulty index, the easier is the item. Classifying items as "good" depends upon the purpose of the test to be developed and the type of items (Kaplan

& Saccuzzo, 1982). In relation to the purpose of a test, if a test is assembled to select candidates who will enter a scholarship program, then the more difficult items will be chosen, since the institution with the program intends to support brilliant candidates. On the other hand, if the test is designed for pretest purposes before a program of study is undertaken, then the less difficult items are more appropriate, since instructors want to get a description of the average knowledge of their participants.

In dealing with the type of items, the chance that students correctly guess a true-false item is 50%. Thus, composing a test which consists of true-false items each with a difficulty level of 0.50 would be considered as a bad test. As another example, a multiple-choice test with 5 options has a chance probability that students will answer each item correctly of 0.20. So, using a difficulty level of 0.20 for each item in composing a test would result in a poor test. As a rule of thumb, for a four-option multiple-choice test format, a good average difficulty level is 0.63. Kaplan and Saccuzzo (1982) explain this

by subtracting 25% (chance level) from 100%, and then dividing by two to find the half-way point. This value is then increased by adding the chance value (0.25):

$$\frac{(1.00 - 0.25)}{2} + 0.25 = 0.63$$

In selecting items to assemble tests, Kaplan and Saccuzzo (1982) suggest that the difficulty level of the items should range from 0.30 to 0.70 to get the maximum information about students' achievement.

#### Item Discrimination

The item discrimination index provides information about whether students who perform well on an item also perform well on the whole test (Kaplan & Saccuzzo, 1982). More specifically, Gronlund (1985) explains that the discriminating power of achievement tests refers to the degree to which tests discriminate between students with high and low achievement. There are several ways to compute item discrimination. The point-biserial method is most relevant to the methodology used in the present study.

The point-biserial correlation is especially applicable for correlation between two variables, one

of which is dichotomous (the item score: right/wrong) and the other being continuous (the total test score). The point-biserial correlation between item and total score is:

$$r_{p-bis} = \frac{(\bar{Y}_i - \bar{Y})}{S_y} \sqrt{\frac{P_i}{(1 - P_i)}}$$

where,  $r_{p-bis}$  = the point-biserial (p-bis) correlation or index of discriminability

$\bar{Y}_i$  = the mean score on the group of students who answered item  $i$  correctly

$\bar{Y}$  = the mean score of all students

$S_y$  = the standard deviation of all students

$P_i$  = the proportion of students who answer the item correctly (Kaplan & Saccuzzo, 1982).

To understand this formula, suppose, for example, 78% of students on a Mathematics test answered Item 15 correctly, and the mean score of students who got Item 15 correct was 57.50. The mean score of the whole mathematics class was 54.50, and the standard deviation on the test was 10.50. Thus, the item discriminating power for Item 15 is:

$$\begin{aligned} r_{p-bis} &= \frac{57.50 - 54.50}{10.50} \sqrt{\frac{0.78}{1 - 0.78}} \\ &= +0.54 \end{aligned}$$

The interpretation of this index is that Item 15 tends to be answered correctly more often by students who have a high score on the total test than by those who have a low score on the total test.

There are three directions in interpreting item discrimination indexes (a) positive discriminating items, (b) negative discriminating items, and (c) non-discriminating items. Items which have positive point-biserial indexes indicate that students who have high score in total score tend to answer these items correctly. Items which have negative point-biserial indexes tend to be answered correctly more often by students who have low scores than by those who have high scores on the total test. Non-discriminating items show that there is no relationship between item score and test score (Popham, 1981). In determining what numerical value might be categorized as a good discrimination index, Popham (1981, p. 298) suggests the figures in Table 2-1.

Table 2-1  
Evaluation of Discrimination Index

DISCRIMINATION INDEX	ITEM EVALUATION
0.40 and above	Very good item.
0.30 - 0.39	Reasonably good but possibly subject to improvement.
0.20 - 0.29	Marginal items, usually needing and being subject to improvement.
0.19 and below	Poor items to be rejected or improved by revision.

#### Distractor Analysis

Another part of item analysis important in a multiple choice format test, is distractor analysis. By examining the distribution of responses for each option of an item, test developers are able to see how effectively each option operates between students who have mastered the course contents and those who have not. For example, suppose that the distribution of students selecting the five options for a particular item is as follows:

	A	B*	C	D	E	Omit
Upper 15 students	2	5	0	8	0	0
Lower 15 students	4	10	0	0	0	1

\* answer key

Option D looks inviting only to the better students. Almost half of higher scoring students chose D as the correct answer, while no lower scoring students chose D. Option C and E do not work well as distractors, because no students were interested in them. This item would need to be completely revised.

The other way of doing item distractor analysis is by computing a point biserial correlation index for each option. Principally, all incorrect options will have negative point biserial correlation indices, because these incorrect options should attract only the lower group of students.

#### Rasch Model Item Analysis

The Rasch Model is a model of item response theory which has become popular in testing. Item response theory is the newest approach in testing and measurement activities, becoming popular in the late 1970s (Hambleton, 1983). Item response theory is also called latent trait theory or item characteristic curve theory (Hambleton et. al., 1978). There are a variety of item response theory models, and the difference among these is primarily in the number of parameters which are used to explain individual item



characteristics.

Hambleton and Swaminathan (1985) state that there are four models of item response theory: the one-parameter logistic (Rasch Model), the two-parameter logistic, the three-parameter logistic, and the four-parameter logistic.

The one-parameter logistic model in explaining individual ability is based on the assumption that items have equal ability to discriminate and that there is no guessing factor. The two-parameter logistic model assumes that the probability of the individual to answer items correctly is influenced by the item difficulty and the item discrimination. The three-parameter logistic model actually comes from the two-parameter logistic model, but in analyzing the items, this model involves guessing as an influencing factor. The four-parameter model, besides considering those three factors mentioned, also takes into account the high-ability examinees who do not always answer the test items correctly (see Hambleton & Swaminathan, 1985).

Compared to other previous methods of item analysis, such as classical test theory, Bejar et al.

(1977) state that item response theory has more advantages:

- Invariance of item parameters. This means that regardless of the distribution of ability of the sample on which we happen to estimate the parameters of the model, the parameter estimates will be linearly related to the parameters estimated with some other sample drawn from the same population. However, we cannot assume that invariance will follow the mere fact of applying the model. That is, invariance must be established empirically.
- Invariance of ability parameters. Another advantage of IRT models is that it is possible to compare two persons' ability estimates even though they may have taken different items. This advantage of IRT permits innovative testing applications such as tailored testing (e.g., Lord, 1980a, Ch. 10)
- A third advantage of IRT is the availability of local measure of precision. That is, unlike classical test theory, which characterizes the impression of test scores by a single value, the standard error of measurement, IRT characterizes precision of measurement by means of a function known as the information, which indexes how precise different scores are. This is a more useful way of characterizing precision of measurement since it specifically allows the fact that precision may be higher for certain values of ability (pp. 3-4).

An important characteristic of item response theory model is that it states explicitly the relationship between the probability of answering an item correctly and the students' ability or level of achievement

(Bejar et. al., 1977, p. 3). They state that an obstacle lies in the way of applying item response models: parameter estimation (Bejar et. al., 1977, p.3).

In order to estimate item difficulty precisely, a requirement is a large sample group (Bejar et. al., 1977, p.3). To make item response theory more usable, Hambleton (1985) adds some considerations. These considerations are as follows:

1. should the model be chosen so that it fits the data well or should the data be edited so the data fits the models desired?
2. the quality of the data,
3. the available resources,
4. the choice of estimation procedure,
5. the availability of computer programs,
6. the assessment of model fit (pp. 307-309).

The Rasch Model, which is also called the one-parameter model of item response theory, was chosen for item analysis in this project because it is a simple model. O'Brien and Tohn (1984) state that only the Rasch model might be applied to a small sample group. In describing the student's ability, it depends upon only one parameter, called the item difficulty. The term 'ability' could be manifested in the form of

achievement for achievement tests, or aptitude for aptitude tests. Wisniewsky (1986) puts forward the advantages of the Rasch Model:

The one model which attempts to explain a response to an item by an individual in the simplest form and by the most elegant design is the Rasch model. It depends on only two parameters-person ability and item difficulty (p.3).

With regard to this quotation, the person ability parameter is a function of only one-parameter item difficulty. These two parameters have the same scale. Hulin, Drasgow, and Parsons (1983) suggest:

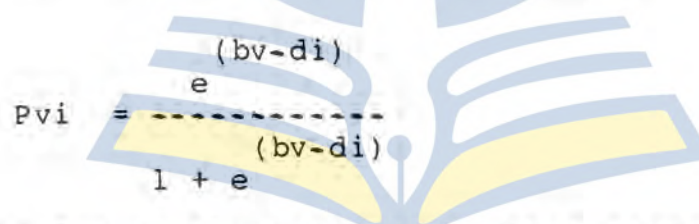
This model is perhaps most useful when a researcher has carefully pretested a set of items that were written in a format that minimizes guessing (p.38).

The Rasch Model requires the assumptions that the result of testing is not influenced by guessing, and that the discriminating power of each test item should be equal. To meet these assumptions, preliminary analysis of test items using classical item theory is the best solution.

The theory assumes one ability parameter  $\theta_v$  for each person  $v$ , and one difficulty parameter  $d_i$  for each

item  $i$ . Because we do not know the actual item difficulty and person ability, they both are called latent variables. In the Rasch Model, both parameters in combination determine the probability of person  $v$  correctly answering item  $i$  (Wright, 1977).

The difference between  $b_v$  and  $d_i$  ( $b_v - d_i$ ) determines the probability of what is assumed to occur if person  $v$  with ability  $b$  faces an item  $i$  with difficulty  $d$ . Both parameter values vary from minus infinity to plus infinity. However, the probability of correct response must be between zero and one, so, to solve this problem, the difference ( $b_v - d_i$ ) should be applied as the exponent of a base  $e$  ( $b_v - d_i$ ). The Rasch formula of the probability for a right answer is:

$$P_{vi} = \frac{e^{(b_v - d_i)}}{1 + e^{(b_v - d_i)}}$$


where,  $P_{vi}$  is the probability that person  $v$  correctly answers item  $i$

Figure 2-1 depicts the probability that person  $v$  will correctly answer item  $i$  depends upon the difference between person's ability  $b$  and item difficulty  $d$ . If ( $b_v - d_i$ ) is equal to zero, the chance

to have right answer is 0.50.

If person  $v$  has more latent ability than the item has latent difficulty, the chance of success is greater than 0.50. In contrast, if the item  $i$  has more of the latent difficulty than the individual has of the latent ability, the probability of success is less than 0.50.

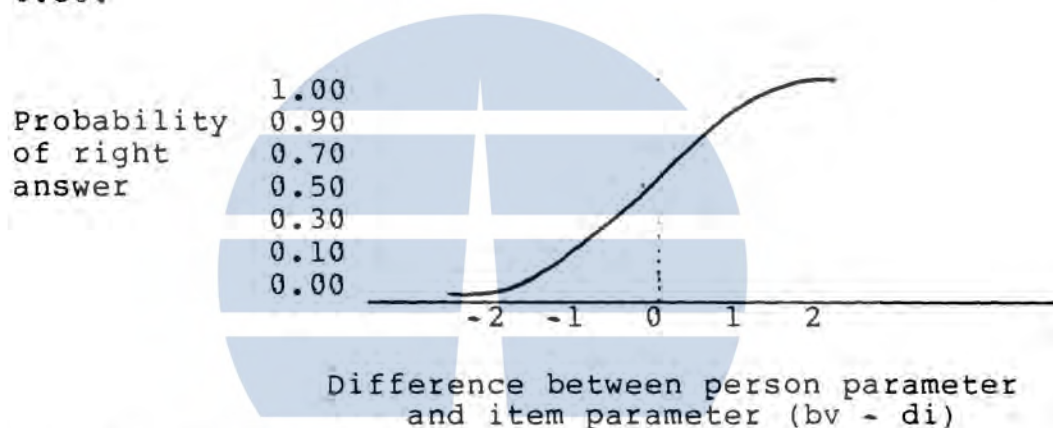


Figure 2-1 The Rasch Model Characteristic Curve  
(From: Wright, 1977)

### Estimating Item Difficulty and Person Ability

To make use of the Rasch Model in practice is to find the difficulty estimate of items and the ability estimate of students, because the purpose of developing tests is to estimate the student ability (trait) (Hambleton & Swaminathan, 1985). The estimation of the item parameter (item difficulty) relates to the

proportion of students who do not answer items correctly. The estimation of the person parameter (student's ability) relates to the students' total test score (Willmott & Fowles, 1974). Wright (1977) provides an example of hand calculation for computing the estimation of the difficulty level and the ability level. But, because computing technology has increased tremendously recently, the process of estimating both the difficulty and the ability parameter has changed from hand to computer. Thus, more teachers can take advantage of item banks than before. One of the first programs created to solve the calibration problem using the Rasch Model was BICAL, developed by Wright and Panchapakesan (Hambleton & Swaminathan, 1985). This program was designed for a mainframe computer. More recently, the Assessment System Corporation developed the program RASCAL for the IBM Microcomputer (Assessment System Corporation, 1986).

### Item Bank Theory

#### Definition of Item Banks

Synonymous for item banks are: question banks, item pools, item collections, item reservoirs, and item libraries (Millman & Arter, 1984). These names

indicate that the item bank can be used in maintaining items, in generating tests and in editing items. Choppin (1978) and Mead (1981) differentiate the meaning of all these terms. They emphasize that the term "item bank" involves more than a collection of items which is implied in the terms "item pool", "item collection", or "item library". Item banks are a collection of items calibrated onto a common scale. Hambleton and Swaminathan (1985) emphasize that an item bank consists of a large amount of items which are matched to objectives, skills or tasks. Thus, when test developers want to develop tests, they can assemble tests easily and accurately. Millman and Arter (1984) suggest that items in a good item bank system should be easily retrieved. Items in the bank must be indexed and structured. This condition can be met when the computer does all the work.

In my opinion, three principles in developing item banks derive from the preceding discussion. Items should be calibrated, items should be relevant to the curriculum, and items should be easy to retrieve. All are requirements for developing an item bank system, because the item bank must function to support student



evaluation. Test items should provide accurate measures of student abilities and the bank should be easy to use.

The latent trait approach to item calibration is attractive, because it is said to be "item-free." This means that different tests can be used to compare individuals. So, teachers can draw any items from the bank to compose a test, say test A. Then another teacher can pull out other items from this bank to assemble another test, say test B. Then they administer these two tests to two different groups of students. The scores of individuals in the two groups which used two different tests can be compared to each other, since each item stored in the bank has been calibrated onto the same scale.

Mead (1981) classifies item calibrations in three ways:

1. One test-form calibration,
2. Two test-form calibration, and
3. Several test-form calibration.

In one test-form calibration, the item difficulties are computed from the performance of all students on all items. In the two and three test-form calibration,

two steps should be done: the students complete the two tests separately and the mean difficulty each is set to zero. By examining scores on common items which have been determined before assembling the tests, a distance between the two tests can be determined. Adjustments can then be made to ensure that all items are calibrated on the same difficulty scale (Wright, 1978).

#### Item Bank Types

Arter and Estes (1985) classify item banking systems into six types:

1. File of Tests,
2. Card file of items,
3. Item stored on computer (use existing word processing programmes),
4. Item information stored on computer (use existing data base management with possible development of some software),
5. Both items and item information on computer; features limited (use micro computer-assisted packaged software), and
6. Sophisticated computerized systems both item and item information on computer; features extensive (use main frame computers) (pp. 44-45).

Type 4 and type 6 are frequently and widely used in instructional activities (Arter & Estes, 1985), and much software is available for item banking (see Dennis et. al., 1985).

### Item Bank Functions

The item bank basically performs three functions: (1) test development, (2) test production, and (3) item maintenance (Burke, Kaufman, & Webb, 1985). These three functions can not be separated from one another, otherwise the bank will not optimally support student evaluation.

#### Test Development

The process in developing tests involves several activities. Hambleton and Swaminathan (1985) delineate the steps of the test development process:

- a. Preparation of test specifications;
- b. Preparation of the item pool;
- c. Field testing the items;
- d. Selection of test items;
- e. Compilation of norms (for norm-referenced tests);
- f. Specification of cutoff scores (for criterion-referenced tests);
- g. Reliability studies;
- h. Validity studies;
- i. Final Test Production (p. 226).

According to Hambleton and Swaminathan, the critical steps are in steps c, d, and g. In step c, the field testing, the test items are tried out to get information about the item characteristics. The representativeness of the subjects in the tryout in the

population is a big issue. The items should be tried out carefully in order to get good items which can differentiate between students on the basis of ability. According to Hambleton and Swaminathan (1985) to be representative of a population, a sample should have at least 300 students. In step d, the selection of test items, the items should measure the abilities required in the curriculum description. The items which are pulled out from the item bank should be related to the curriculum objectives. Finally, in step g, the reliability study is aimed to see if the tests measure the students consistently.

#### Test Production

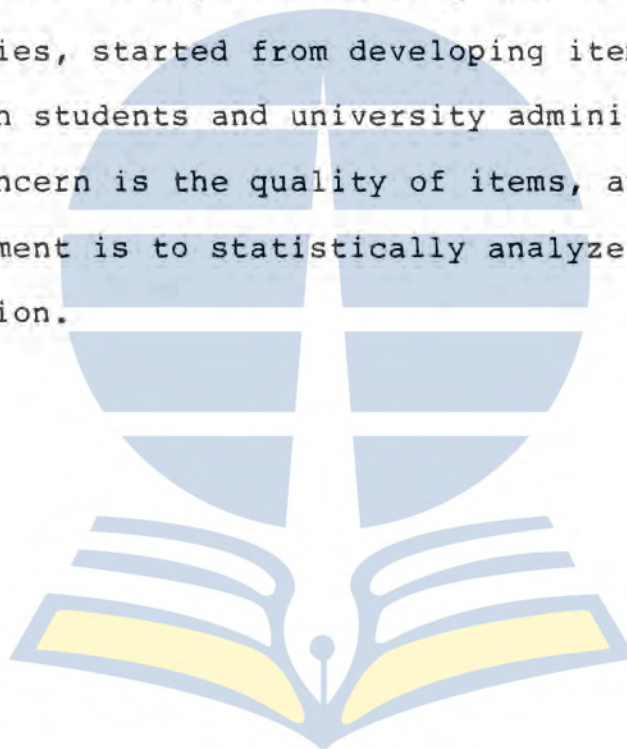
This activity includes laying out test booklets and then printing them. The concern of the item bank management in this function is the format of test books and the instruction to the students, so the students do not incorrectly interpret the tests.

#### Item Maintenance

The three activities in maintaining an item bank are: (a) adding items, (b) updating old items, and (c) dropping items (Mead, 1981). Adding items to the bank can be done at any time depending upon the needs of the

bank. The concern is that those items which are stored should be scaled in order that the items in the bank are calibrated onto a common scale. The item bank sometimes needs to be updated, because the contents are not relevant to the curriculum. If the items can not be edited, then the items are just dropped.

In conclusion, item banking covers all examination activities, started from developing items to reporting for both students and university administration. The main concern is the quality of items, and the basic requirement is to statistically analyze its item collection.



### CHAPTER III

#### METHODOLOGY

The purpose of this project is to develop an item banking system for Universitas Terbuka. Two courses, English and Mathematics, within the Diploma II program at the Faculty of Education will now be used to demonstrate the procedure. The main activities are item analyses and calibration. In this chapter, I discuss the ITEMAN and RASCAL programs. Then, I carry out item analyses and item calibration.

#### ITEMAN and RASCAL

Both classical item analysis and Rasch procedures are already programmed for the IBM microcomputer. These two software programs are ITEMAN (ITEM analysis) for classical item analysis, and RASCAL (RASch CALibration) for item analysis using the Rasch Model (Assessment System Corporation, 1986).

The ITEMAN and the RASCAL programs are parts and subsystems of the MicroCAT Testing System. ITEMAN is appropriate in analyzing items for their difficulty index, discrimination power, and item distractors. RASCAL provides items analysis using the Rasch Model,

and has two scales: the ability and the difficulty scale. If the ability scale is determined to be a standard then the mean of the ability distribution of the sample is set to zero and the variance is set to one. If the difficulty is decided to be a standard then the mean of the difficulty distribution is set to zero and the variance is set to one. Items which can be analyzed may be dichotomous, as in a true-false format, or multipoint items such as multiple-choice formats and attitude scales.

#### Preparing Data Files

The data file consists of five elements. First is a control line which describes the data. Second is a list of correct responses. Third is a list of the numbers of options for each item. Forth is a list categorizing which items are to be analyzed, and fifth is the sample data. Figure 3-1 describes an hypothetical data layout for an item analysis.

In the first line, columns 1-3 describe the number of items which will be analyzed ( maximum 250), column 4, 6 and 8 are blank, column 5 is alphanumeric code for omitted responses, column 7 is alphanumeric code for items not reached by testee, and columns 9-10 show the

number of characters of data identification (maximum characters are 80).

```

20 0 N 5                control line
ACCBDACBADDDBACCABAD  key response
44444444444444444444  options
11111111111111111111  grouping
0001 ACCDDABBADDDBACBAABAD record no.1
0002 ACCBDACBADADDBACBAAN record no.2
0003 ACCBDACBADDDBACCABAD record no.3
0004 ABCBDCCAABDDDBA00ABAC record no.4
.....
....
....
0100 ABCDACCACBDDDBACBABAD record no.100

```

Figure 3-1 An Example of Data Layout

Data for ITEMAN and RASCAL should be formatted in ASCII files. Thus, the data must be written using a word processor or text editor which can produce files in ASCII codes. ITEMAN and RASCAL can analyze data for up to 30 000 testees (Assessment System Corporation, 1986).

### Interpreting the Analysis

#### ITEMAN

ITEMAN produces five item statistics for dichotomously scored items: (a) the sequence number, (b) the item scale, (c) the proportion of correct



responses, (d) the biserial correlation between correct responses to the item and total score test, and (e) the corresponding point-biserial correlation. Figure 3-2 shows an example of analysis that reports item characteristics and alternative statistics.

MicroDAT (tm) Testing System  
Copyright (c) 1982, 1984, 1986 by Assessment Systems Corporation  
Item and Test Analysis Program -- ITEMAN Version 2.0

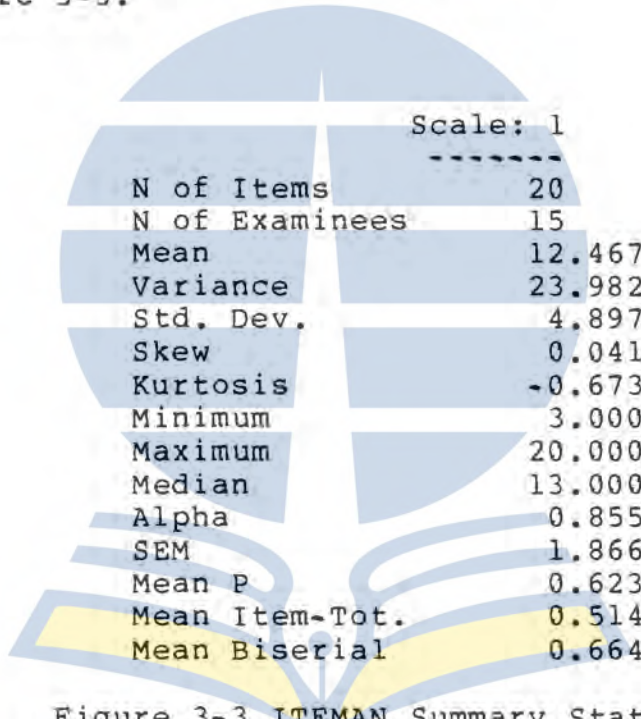
Item analysis for data from file medi.prn Page 1

Item Statistics					Alternative Statistics				
Seq. No.	Scale	Prop. Correct	Biser. Biser.	Point Biser.	Alt. Endorsing	Prop. Biser.	Point Biser.	Key	
1	1-1	0.533	0.968	0.771	A	0.533	0.968	0.771	*
					B	0.200	-0.408	-0.286	
					C	0.067	-0.365	-0.189	
					D	0.133	-0.438	-0.278	
					Other	0.067	-0.997	-0.517	
2	1-2	0.533	0.762	0.608	A	0.067	-0.365	-0.189	
					B	0.267	-0.243	-0.181	
					C	0.533	0.762	0.608	*
					D	0.067	-0.365	-0.189	
					Other	0.067	-0.997	-0.517	
3	1-3	0.800	0.797	0.558	A	0.067	-0.365	-0.189	
					B	0.067	-0.365	-0.189	
					C	0.800	0.797	0.558	*
					D	0.000	-9.000	-9.000	
					Other	0.067	-0.997	-0.517	

**Figure 3-2** The Item Statistics and the Alternative Statistics

The item difficulty is the proportion of the students who answer the item correctly. The biserial and the point-biserial indices can be used as item discrimination indices.

In the summary statistics, the ITEMAN program reports 16 statistics related to test scores, as shown in Figure 3-3.



Scale: 1

N of Items	20
N of Examinees	15
Mean	12.467
Variance	23.982
Std. Dev.	4.897
Skew	0.041
Kurtosis	-0.673
Minimum	3.000
Maximum	20.000
Median	13.000
Alpha	0.855
SEM	1.866
Mean P	0.623
Mean Item-Tot.	0.514
Mean Biserial	0.664

Figure 3-3 ITEMAN Summary Statistics

### RASCAL

Figure 3-4 shows the results of an analysis of 20 hypothetical items using RASCAL when the difficulty is standardized with the mean of zero.

MicroCAT (tm) Testing System  
 Copyright (c) 1982, 1984, 1986 by Assessment Systems Corporation

Rasch Model Item Calibration Program -- RASCAL Version 1.0

Final Parameter Estimates for Data from File medi.prn

Item	Difficulty	Chi Sq.	df
1	0.505	3.976	2
2	0.505	0.785	2
3	-1.077	1.309	2
4	0.505	0.785	2
5	-0.233	0.603	2
6	-0.628	1.655	2
7	-0.628	0.136	2
8	0.505	0.785	2
9	0.138	1.741	2
10	1.317	1.942	2
11	-0.628	0.136	2
12	-0.233	0.594	2
13	0.138	2.011	2
14	-0.233	0.603	2
15	-0.233	3.918	2
16	0.138	2.011	2
17	-0.233	7.277	2
18	0.505	2.852	2
19	-0.628	9.674	2
20	0.505	2.852	2

Figure 3-4 RASCAL Output when the Difficulty Scale is selected

The difficulty level estimation provided by the Rasch model differs from the classical item analysis. The difficulty level in Rasch procedure refers to "the point on the ability scale at which an examinee has 50%

chance of answering the item correctly" (Rentsz & Rentsz, 1974, p. 2). The difficulty estimate ranges from infinite positive to infinite negative. However, in reality it ranges from +4 logit to -4 logit. A negative value for an item indicates an easy item (Robitaille & O'Shea, 1983).

The chi square is a measure of the degree of fit of the item to the model. In this model, no critical values are suggested, but items showing marked deviation in chi square are suspect. For example, in Figure 3-4, Item 19 could be suspected as not fitting the model because of the high chi square value.

#### Item Analysis Report

In this study, the tests were chosen from the final examinations in 1986. Take-home assignments were not used, because items on these tests did not go through the test administration procedure. The item collections at Universitas Terbuka are in multiple-choice format, with four options.

For item analysis using Rasch Model, *ten* items were selected from December 1986 examination and were also used in May 1987. The purpose of using these 10 common items was to adjust the difficulty level of the

items used in December 1986 and in May 1987. In the item selection stage for the May 1987 tests, I was not permitted to select items. The Faculty members only had the authority in selecting the 10 items to be reused in May 1987.

Analysis of the items takes into account the item difficulty index, the item discrimination index, and the distribution of responses across item distractors. Furthermore, the Rasch Model item analysis was also employed, because this model provides an excellent way to calibrate items onto a common scale. Two topics are discussed here, and they describe the chronological stages in item selection: collecting student score records, and analyzing items.

#### Collecting Student Scores

The student score data from the English and Mathematics examinations was sorted and captured from the computer mainframe at Universitas Terbuka. These scores were then saved into a 5.25 inch floppy disk for item analyses using an IBM PC.

There were 45 students who took the English examination and 166 students who took the Mathematics examination in December 1986. In May 1987, the number

of students who took these two exams increased to 73 students in the English course, and 323 students in the Mathematics course.

Both the English and Mathematics test booklets contained 80 multiple-choice format items in December 1986. In May 1987, the number of items in the mathematics booklet was reduced to 45, but the number of items in the English booklet remained constant. Before analyzing items, some items in May 1987 were deleted. Item 25 in the English examination was dropped because the answer was incomplete. In the Mathematics examination, Items 21, 23, 24, 36, and 45 were deleted because there were no tables attached to each item.

#### Analyzing Items

Item analysis consisted of three main activities. The first was entering the raw data into the format required by the MicroCAT Testing System using the Symphony word-processing program. Appendix B provides a sample of the raw scores. The second was running ITEMAN and RASCAL programmes. The third was doing the item analysis using classical item analysis followed by the Rasch model. The following is a further

description of both analyses.

#### Classical Item Analysis

The purpose of this analysis is to select items which will be stored in the bank. The main concerns were the discriminating power of each item and the distribution of responses across the incorrect options.

The difficulty level was not considered in this stage, because I anticipate that in the future the bank will not only provide items for assembling achievement tests as such, but it will provide items for pretesting students as well.

The complete sequence of item analysis was as follows:

- a. A classical item analysis was performed using the ITEMAN program. The classical item analysis provided information about the item difficulty index and the point-biserial correlation index for each option. The sample preliminary analyses are reported in Appendix C.
- b. Based upon the preliminary analysis, I selected items using the following criteria:
  1. The point-biserial correlation index for the correct answer is 0.20 or higher. Based on

Popham's (1983) suggestion, items which have a discrimination index between 0.20 and 0.29 are considered as marginal items. These items could be stored in the bank with some revision when the faculty members plan to reuse them.

2. All incorrect options should be chosen by some students, and there should be a negative correlation index with the total score. However, I did not apply this criterion strictly. I included items for which only one of the distractors had a positive value. I set the maximum limit at 0.05. Table 3-1 contains a summary of items in English and Mathematics which have borderline distractors.

Table 3-1  
Summary of Items Having Borderline Distractors

Item Number	Option	Alternative Statistics	Key
-------------	--------	------------------------	-----

1. English (December 1986)

1	A	0.042	C
26	C	0.004	B
38	B	0.050	A
60	B	0.026	D
65	D	0.050	A



Table 3-1 (Continued)  
 Summary of Items Having Borderline Distractors

Item Number	Option	Alternative Statistics	Key
<u>2. English (May 1987)</u>			
10	A	0.009	B
32	A	0.025	C
46	B	0.009	C
58	C	0.040	B
67	D	0.038	A
<u>3. Mathematics (December 1986)</u>			
11	D	0.034	A
14	A	0.011	B
17	C	0.041	B
19	D	0.044	A
20	D	0.000	B
36	D	0.010	A
39	C	0.037	B
76	B	0.032	C
<u>4. Mathematics (May 1987)</u>			
4	C	0.045	A
19	D	0.048	B
31	A	0.038	C
38	A	0.006	D

- c. The rest of the items were deleted because they had poor discrimination power and ineffective distractors. For example, item 7 in English (December 1986) was deleted because nobody chose option D. The other items in English (December 1986) which had bad distractors were items 10, 12, and 39.

Based on this analysis, 30 items in the December 1986 test and 43 items in the May 1987 test for English were selected. 29 items in December 1986 and 28 items in May 1987 for Mathematics were also selected. These selected items for those four tests are provided in Figure 3-5.

1. English in December 1986 (30 items)

-----  
 1, 6, 11, 13, 16, 17, 18, 19, 26, 27, 32, 33, 35, 37,  
 38, 42, 44, 49, 50, 52, 58, 59, 60, 61, 62, 63, 65, 69,  
 70, and 78

2. English in May 1987 (43 items)

-----  
 2, 3, 6, 7, 8, 9, 10, 11, 13, 15, 16, 17, 19, 24, 30,  
 31, 32, 33, 35, 36, 38, 39, 40, 42, 43, 46, 48, 49, 50,  
 51, 53, 54, 55, 57, 58, 61, 62, 67, 69, 73, 76, 77, and  
 78.

3. Mathematics in December 1986 (30 items)

-----  
 1, 2, 3, 5, 6, 7, 8, 9, 11, 12, 14, 15, 17, 19, 20, 21,  
 23, 25, 26, 27, 28, 30, 36, 38, 39, 48, 59, 69, 71, and  
 76.

4. Mathematics in May 1987 (28 items)

-----  
 1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,  
 18, 19, 20, 25, 27, 28, 31, 35, 38, 39, 42, and 44.

Figure 3-5 The Selected Items Based on Classical Item  
 Analysis

d. The selected items were reanalyzed using the ITEMAN. The purpose of reanalyzing is to find established indexes. The results of the second analyses using the classical item analysis are reported in Appendix D.

#### Rasch Model Item Analysis

The computer program for the calibration is the RASCAL program developed by The Assessment System Corporation. The complete results of item analysis are reported in Appendix E.

The next activity was to analyze the 10 items which were used in both examination periods. Table 3-3 and Table 3-4 list items which were used in the two examination periods and the result of the Rasch analysis.

In the English tests, only four items (marked by \*) were valid for the equating process. Most items were dropped because they did not conform to the requirement that all options be selected. Perhaps the student samples were not quite large enough; 47 students in the December 1986 and 73 students in the May 1987.

In the Mathematics tests, item number 21 was dropped, because when it was printed in the booklet,

Table 3-2  
English Items used in December 1986  
and in May 1987

December 1986		May 1987	
Item No.	Rasch Difficulty	Item No.	Rasch Difficulty
7	(deleted)	10	-0.213
6	1.104	11	1.594 *
12	(deleted)	17	-0.530
10	(deleted)	20	(deleted)
17	-0.011	21	(deleted)
13	-0.122	32	0.156 *
39	(deleted)	48	-0.530
44	-0.905	51	-0.448 *
32	-0.454	52	(deleted)
49	-1.023	57	-0.290 *

\*) items used for calibrating process

Table 3-3  
Mathematics Items used in December 1986  
and in May 1987

December 1986		May 1987	
Item No.	Rasch Difficulty	Item No.	Rasch Difficulty
5	-0.703	2	-0.669 *
6	-2.434	3	-1.617 *
7	-1.107	5	-1.525 *
15	-0.734	7	-0.637 *
12	-1.419	8	-1.952 *
23	-1.419	12	-1.395 *
26	-0.201	14	0.182 *
28	-0.145	15	0.560 *
38	1.236	21	(deleted)
48	-0.519	25	-1.816 *

\*) items used for calibrating process

it did not contain a necessary table.

The purpose of this analysis is to calibrate items onto a common difficulty scale. When calibrating items using the Rasch Model, it necessary to assume that the ability of two groups, in December 1986 and in May 1987, are the same. Table 3-2 is the summary of students's score based upon the 4 common items in English and 9 common items in Mathematics.

Table 3-4  
Summary of Students' Score on Common Items

	ENGLISH		MATHEMATICS	
	1986	1987	1986	1987
MEAN:	2.22	2.22	6.48	6.37
SD:	1.43	1.38	2.03	1.85
N:	45	73	166	323
MAX:	4	4	9	9
MIN:	0	0	0	0

On the basis of the results presented in Table 3-2, I conclude that the ability of both group in English and in Mathematics have the same ability.

#### Adjusting Item Difficulties to a Common Scale

The following are the steps in the process for

scaling item difficulties in each subject area.

- a. The common items used in the two examination periods were compared to determine their difference in difficulty.
- b. The mean of the differences in item difficulty was determined.
- c. The difficulty level of all items in each testing period were adjusted to a common scale.
  1. For the items which were used twice, find the adjusted difficulty level by averaging the scores from December 1986 and from May 1987. Table 3-5 contains the summary of mean differences and the adjusted difficulties. For the English test-items, only four valid items were used in the adjustment process.
  2. The remaining items were adjusted as follows: (a) For English test items on December 1986 exam, adjust all difficulty level up by 0.23 ( $=0.45/2$ ) and adjust all difficulty level down by 0.23 for test items in May 1987 examination period, because items in May 1987 were more difficult

Table 3-5  
Summary of Computing the Mean Difference  
and the Adjusting Process

ENGLISH					
DEC '86	MAY '87		Difference	Adjusted	
Item No.	Diffic.	Item No.	Diffic.	Between	Difficulty
			MAY-DEC		
6	1.14	11	1.59	0.46	1.37
13	-0.08	32	0.16	0.24	0.04
44	-0.86	51	-0.45	0.42	-0.66
49	-1.98	57	-0.29	0.69	-0.64
			Total :	1.80	
			mean difference :	$1.80/4 = 0.45$	
MATHEMATICS					
DEC '86	MAY '87		Difference	Adjusted	
Item No.	Diffic.	Item No.	Diffic	Between	Difficulty
			DEC-MAY		
5	-0.68	2	-0.68	0.01	-0.68
6	-2.41	3	-1.63	-0.78	-2.02
7	-1.08	5	-1.54	0.45	-1.31
15	-0.71	7	-0.65	-0.06	0.68
12	-1.39	8	-1.97	0.57	-1.68
23	-1.39	12	-1.41	0.01	-1.40
26	-0.18	14	0.17	-0.34	-0.01
28	-0.12	15	0.54	-0.66	0.21
48	-0.49	25	-1.83	1.33	-1.16
			Total:	0.54	
			mean difference:	$0.54/9 = 0.06$	

than in December 1986. (b) For Mathematics test items in December 1986 exam, adjust all the difficulty level down by 0.03 ( $= 0.06/2$ ) and adjust all difficulty level up by 0.03 for test items in May 1987 examination period, because items in December 1986 were more difficult than in May 1987. Table 3-6 contains the result of the adjustment process.

Table 3-6  
Adjusted Difficulties

1. English Test

DECEMBER 1986			MAY 1987		
Item No.	Diff.	Adjusted	Item No.	Diff.	Adjusted
1	-0,64	-0,41	2	-1,29	-1,52
* 6	1,14	1,37	3	-0,29	-0,52
11	1,14	1,37	6	-1,65	-1,88
* 13	-0,08	0,04	7	0,37	0,14
16	-0,64	-0,41	8	0,16	-0,07
17	0,03	0,26	9	-0,61	-0,84
18	-0,64	-0,41	10	-0,21	-0,44
19	0,37	0,60	* 11	1,59	1,37
26	1,00	1,23	13	-0,14	-0,37
27	-0,53	-0,30	15	-1,94	-2,17
32	-0,41	-0,18	16	0,23	0,00
33	-1,10	-0,87	17	-0,53	-0,76
35	0,14	0,37	19	-0,21	-0,44
37	0,86	1,09	24	-0,14	-0,37



Table 3-6 (Continued)  
Adjusted Difficulties

DECEMBER 1986			MAY 1987		
Item No.	Diff.	Adjusted	Item No.	Diff.	Adjusted
38	-2,38	-2,15	30	0,72	0,49
42	0,74	0,97	31	0,79	0,56
* 44	-0,86	-0,66	* 32	0,16	0,04
* 49	-0,98	-0,64	33	0,86	0,63
50	-0,08	0,15	35	-0,61	-0,84
52	0,14	0,37	36	-0,06	-0,29
58	0,03	0,26	38	-0,14	-0,37
59	-0,41	-0,18	39	-0,29	-0,52
60	0,37	0,60	40	-0,21	-0,44
61	1,97	2,20	42	0,23	0,00
62	-0,08	0,15	43	-1,40	-1,63
63	0,25	0,48	46	0,79	0,56
65	0,03	0,26	48	-0,53	-0,76
69	0,14	0,37	49	-0,45	-0,68
70	0,14	0,37	50	-0,98	-1,21
78	0,37	0,60	* 51	-0,45	-0,66
			53	0,37	0,14
			54	0,79	0,56
			55	-0,14	-0,37
			* 57	-0,29	-0,64
			58	-0,06	-0,29
			61	0,23	0,00
			62	0,08	-0,15
			67	1,76	1,53
			69	0,37	0,14
			73	0,37	0,14
			76	0,86	0,63
			77	1,00	0,77
			78	0,93	0,70
Total:	0,01	6,89	Total:	0,00	-9,88
Mean of Dec:		0,23	Mean of May:		-0,23
Standard Dev:		0,81	Standrd Dev:		0,76

## 2. Mathematics Test

DECEMBER 1986			MAY 1987		
Item No.	Diff.	Adjusted	Item No.	Diff.	Adjusted
1	0,49	0,46	1	-0,14	-0,11
2	0,02	-0,01	* 2	-0,68	-0,68
3	0,58	0,55	* 3	-1,63	-2,02
* 5	-0,68	-0,68	4	2,16	2,19
* 6	-2,41	-2,02	* 5	-1,54	-1,31
* 7	-1,08	-1,31	* 7	-0,65	-0,68
8	-0,62	-0,65	* 8	-1,97	-1,68
9	-0,23	-0,26	9	-0,07	-0,04
11	1,05	1,02	10	-0,30	-0,27
* 12	-1,39	-1,68	11	-0,64	-0,61
14	0,35	0,32	* 12	-1,41	-1,40
* 15	-0,71	-0,68	* 14	0,17	-0,01
17	1,17	1,14	* 15	0,54	0,21
19	0,32	0,29	16	1,05	0,35
20	0,80	0,77	17	0,34	0,37
21	-1,27	-1,30	18	-0,62	-0,59
* 23	-1,39	-1,40	19	0,99	1,02
25	-0,94	-0,97	20	0,41	0,44
* 26	-0,18	-0,01	* 25	-1,83	-1,16
27	-0,06	-0,09	27	0,89	-1,17
* 28	-0,12	0,21	28	1,02	1,05
30	1,24	1,21	31	1,22	1,25
36	1,44	1,41	35	0,69	0,72
38	1,27	1,24	38	0,01	0,04
39	0,89	0,86	39	-1,09	-1,06
* 48	-0,49	-1,16	41	0,62	0,65
59	-0,23	-0,26	42	1,31	1,34
71	1,24	1,21	44	1,15	1,18
76	0,95	0,92			
Total:	0,00	-0,88	Total:	0,00	-1,98
Mean of Dec:		-0,03	Mean of May:		-0,07
Standard Dev:		0,98	Standrd Dev:		1,02

These items with their difficulty estimates,  
computed by RASCAL and their discrimination indices

computed by ITEMAN were stored into the bank to assemble tests.



## CHAPTER IV

### PROPOSED ITEM BANK SYSTEM

The purpose of developing an item bank system at Universitas Terbuka is to develop, and to maintain items which will be used to assemble tests for final examinations. These tests have an important role in evaluating the student achievement, because this is the only device to measure the students' performance after taking distance courses.

In this chapter, I discuss the condition of Universitas Terbuka's item bank and outline a proposal for an item bank system. Following Millman and Arter (1984), I will address four crucial topics: (a) item development, (b) the use of item bank, (c) test assembling, and (d) test administration.

#### Item Development

To develop items intended to support the item bank system at Universitas Terbuka, I will use the existing item development procedure already implemented at Universitas Terbuka, with some additional requirements. The general procedure for item development is as follows:

1. Develop a matrix of test content based on the curriculum objectives. This activity has been done by the course writers or the faculty members who are responsible for the teaching and learning process. For example, the items for the English course, and the matrix of the content of the examination, were developed by course writers who were responsible for this English course. The layout of the test matrix is described in Figure 4-1.
2. Write items. The item writers will be the faculty members. If there are no competent faculty members then course writers or other competent persons will write items. If the bank has been established, the faculty members will write 10 to 20 items, which relate to the curriculum objectives, to add to the bank collections.
3. Overview items. After the items have been created, they will be reviewed by content experts such as university members from other universities, or the course writers themselves. If some items do not satisfy the review committee, then the item writers rewrite these items, and resubmit them to the committee.

Faculty	: M&NS <sup>1)</sup>	Semester	: ____/XII		
Academic Program	: Ap. Stat <sup>2)</sup>	The Writer	: Mr. _____		
Course	: Math 3	Examination	: FE/TA <sup>3)</sup>		
Ability Level	C1,2	C3	C4,5	C6 <sup>4)</sup>	Sum of
Item Format	ABCDE	ABCDE	ABCDE	<sup>5)</sup> Item	
No. Topic and Instructional Objective					
1. <u>System &amp; Component</u>					
a. _____	?	-	-	-	??
b. _____	?	?	-	-	??
c. _____	-	?	?	-	??
2. <u>Set</u>					
a. _____	?	?	-	-	??
b. _____	-	?	?	-	??
c. _____	-	-	?	?	??
3. _____ etc.					
	20%	40%	30%	10%	100%
<sup>1)</sup> Mathematics and Natural Science <sup>2)</sup> Applied Statistics <sup>3)</sup> FE = Final Examination TA = Take-home Assignment <sup>4)</sup> C1,2 = Knowledge and Comprehension C3 = Application C4,5 = Analysis and Synthesis C6 = Evaluation <sup>5)</sup> A = Completing four options B = Choice of four combination of options C = Case Analysis D = Multiple Completion E = Analysis of Diagram					

Figure 4-1 Example of an Examination Matrix

4. Items which satisfy the committee will be tried out in the field. These tried out by attaching them to the 50 or 60 calibrated items drawn from the bank for testing purposes. These 10 to 20 items, however, will not be involved in student grading. The purpose is to estimate their statistical characteristics.

I am concerned that the faculty members write items rather than having outside test writers do so. The advantages when the faculty members develop items by themselves are (1) they will have experience in developing evaluation tools in their faculty, (2) they will have a responsibility in the evaluation process, and (3) it will save money, because writing items will be part of their routine job.

#### Item Format

Items stored in the bank are objective-tests which use four options in a multiple-choice format. The various formats are as follows:

1. Completing four options. Students have to choose the most appropriate from four responses. For example:

He knows a lot about commercial affairs.  
The correct noun form for commercial  
affairs is ...

- A. commercialization
- B. commerce
- C. commerciality
- D. commercence

2. Choice of combinations of options. Students should:

- (a) choose A, if option 1 and 2 are correct,
- (b) choose B, if option 1 and 3 are correct,
- (c) choose C, if option 2 and 3 are correct, and
- (d) choose D, if all options (1,2, and 3) are correct, For example:

I had been thinking about you when you came means:

- 1. I stopped thinking when you came;
- 2. I was thinking when you came;
- 3. I had started thinking when you came.

3. Matching. Students have to choose one of four options which matches the question, for example:

Direction: Which matches the question tag?  
....., will you?

- A. You and your sister will leave for Bangkok
- B. Please leave me alone
- C. You, your sister and I won't come
- D. You and she will have married



4. Rearrange the words into a correct sentence. For example:

can - many - carry - boat - your - how - passengers - small ?

- A. How many passengers can your small boat carry?  
B. How can your small boat carry passengers many?  
C. How passengers your small boat can carry many?  
D. How can passengers carry your many small boat?
5. Multiple completion. One uncompleted statement is followed by some possible correct answers. Students choose:
- (1) A if options 1 and 2 are correct,
  - (2) B if options 1 and 3 are correct,
  - (3) C if options 2 and 3 are correct, and
  - (4) D if all options are correct.

Items which will be stored in the item bank should be analyzed using both classical and Rasch model item analyses. The classical analysis should be done first to eliminate poor items. Then the selected items should be calibrated using the Rasch Model to get the common difficulty level. This index will be used to

calibrate the item collection onto a common scale.

Items stored in the bank will be classified into three groups based on their difficulty level. Using Robitaille and O'Shea (1983) classification, the groups are: (1) very difficult items, which have p-value between 0% and 29%, (2) difficult items, which have p-value between 30% and 79%, and (3) easy items, which have p-value between 80% and 100%.

#### The Use of the Item Bank

The item bank will be assisted by the computer mainly for item analysis, storage of items, and item generation. Each Faculty will be the regular user of the bank. When the faculty members want to benefit from the bank, they will first make a matrix which describes the curriculum objectives. Second, they will go to the bank or perhaps use a terminal in each faculty to match items and the curriculum objectives items. If they agree with the items selected then they print items to compose tests.

Items stored in the bank use the format for information developed at Universitas Terbuka, which has the following characteristics:

- (1) General characteristics, such as: faculty, study

program, courses, in what semester item used, item code, level of ability (Bloom Taxonomy), difficulty level (easy, average, difficult), item model (matching, choose the appropriate response), key, item writer code, module reference, item reviewer 1, item reviewer 2, associated learning activities, and the instructional objective. (2) Statistical characteristics, such as: the distribution of responses, p-value, and point biserial correlation.

In the use of items, the item bank staff should also provide a manual and training to the user about using the computer, test development, and security. It should be clear that assessment by using a computer is easier than doing everything by hand.

#### Item Maintenance

To maintain the quality of the item bank system, some steps for developing and maintaining items should be considered: (1) training, (2) review materials, and (3) tryout items. In the training step, not only do the item writers need to be trained in order to become qualified item writers but also the item bank users need to be trained in order to be able to utilize the item bank properly and effectively. They must also be

trained to understand the basic ideas of item difficulty and the Rasch model.

In reviewing the items, the first step is evaluating the blueprint, because the test should match the social change. It is not fair if the test blueprints are constant all the time while the social demand is increasing. Second, content analysis should be carried out, because over a period of time, the item content may become familiar to the students. As a result, tests developed from the bank may not reflect the students' ability.

#### Test Assembling

This item bank system project plans to use the computer as the means to store, to select, to edit, and to generate the items. Each item is to be stored using a certain label in order that the items will be easy to select and to retrieve.

In developing achievement tests, the criteria for retrieving items are:

- a. P-value should range between 0.30 to 0.70. This range will optimize the information about the students' ability, because all the items have only 4 options.

- b. Select items which fairly represent the module to be tested.
- c. The total number of items included in one test should be related to the students' ability which is to be measured. For example, for items requiring high mental capability such as in mathematics courses, it is enough to set a test with 50 items for one hour.
- d. Items are arranged starting from the easiest (highest p-value) to the hardest (lowest p-value). By putting items in such way, it will support students in completing their tests.

#### Test Administration, Scoring, and Reporting

To examine the students, the test booklets will be sent by mail to the UPBJJs (regional centers) which are responsible for the test administration and also for the test security. Analysis of the results will be done by the Universitas Terbuka, in Jakarta. There are three activities involved in the item analysis process: (a) analyzing items using classical item analysis, (b) calibrating items using the Rasch Model, and (c) adjusting items to the overall collection.

### System

There are four components in the system: Item Development, Item Selection, Test Administration, and Item Maintenance. Each faculty has the task of developing items. They develop the matrix, organize item writers, review items, and make a decision on the content quality of the items.

In the item selection phase, again the faculty will actively develop a matrix based on the course materials. After they have completely developed the matrix, they select items to compose tests from the terminals which are located in each faculty. Figure 4-1 clarifies the item bank system.

In the Administration stage, the role of the Examination Centre is to coordinate the faculty staff, the examination centre staff, and the UPBJJ to manage test administration. The activities in this stage are printing tests booklets, arranging the test schedule, distributing tests to the UPBJJs, scoring, and reporting. The main task of the examination center is to statistically analyze items, and to code items for item banking purpose.

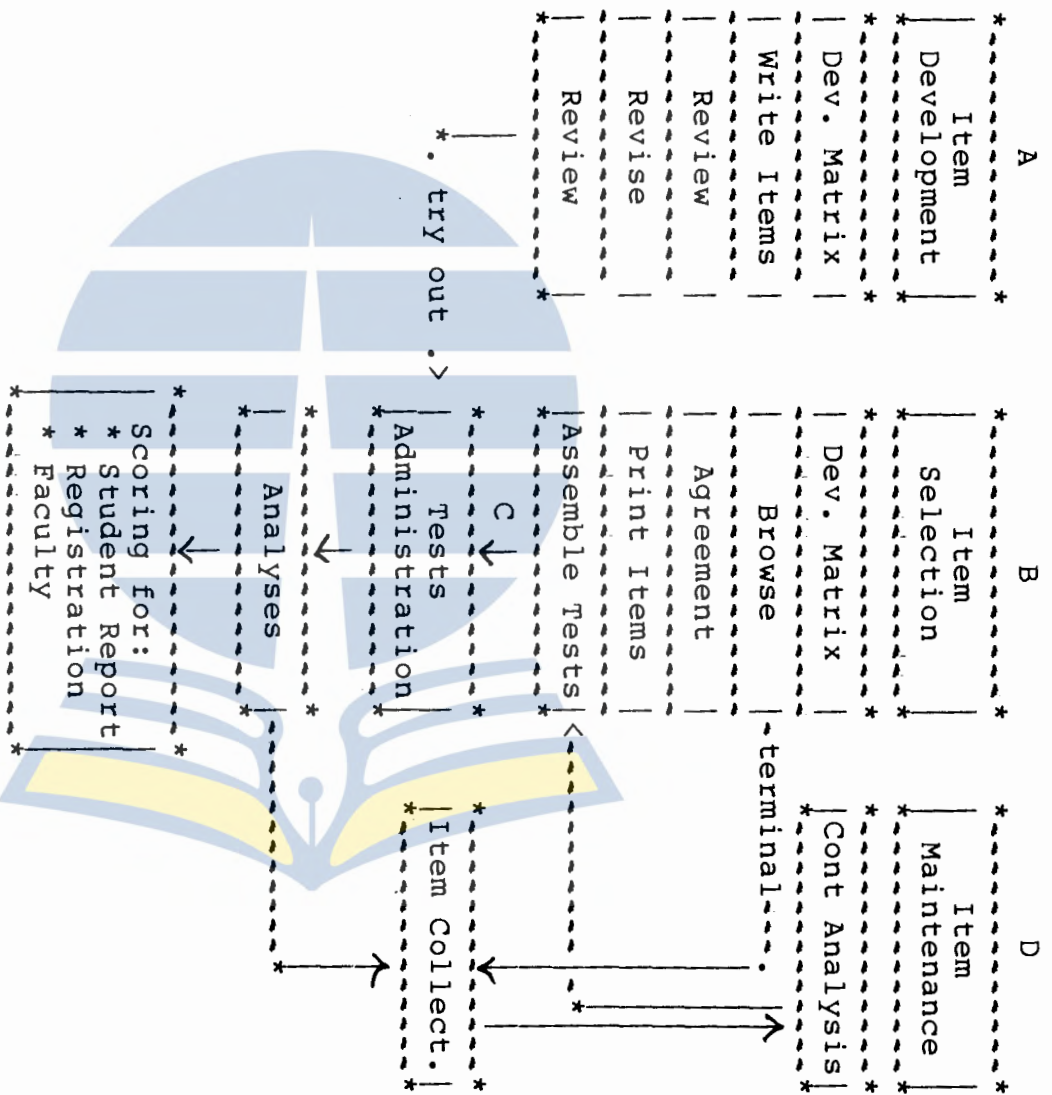


Figure 4-2. An Item Bank System for Universitas Terbuka.

The tasks in the Item Maintenance stage are not as frequent as in the other three stages, because it depends upon the need to make changes. After a period of time when items are frequently drawn from the bank, the contents should be changed but not the objective. The process in this stage, firstly, is to identify items which need to be edited; secondly, to rewrite items; thirdly, to assemble tests; and finally to analyze items and store them into the bank.

In the total system, the Examination Centre's task is to support each faculty in the process of the test items analysis. The Centre also provides manuals and training for item writers and bank users.





CHAPTER V  
DISCUSSION AND RECOMMENDATIONS

Discussion

Item Analysis

The crucial part in developing item banks is to analyze items and calibrate them onto a common scale. Accurate item analysis requires statistical item analysis. Classical item analysis and the Rasch model procedure provide empirical data to select good items and to calibrate items. Items evaluated, for example, by experts as good items do not always measure the student's achievement properly when statistical item analysis is applied. In this study, based upon the statistical item analysis, 38% and 54% of the items were considered to be good items for English in the December 1986 and May 1987 examination periods. In Mathematics, there were 33% and 70% good items in December 1986 and May 1987. That is, about half the test items conformed to the statistical analysis requirements.

One should not, however, conclude on this basis alone that examinations were poor because they contained many poor test items. There were several

constraints in applying statistical item analysis. The first constraint was that statistical item analysis needs large samples, and there were only 46 and 67 students contributing to the item analyses on the English examinations. These small samples could result in inaccurate description of the item characteristics.

The second problem was that the time allotted and the number of test items was not in balance. For example, in the Mathematics examination held in December 1986, many students omitted responses from item 50 to the end of the test. Response patterns indicated that guessing was evident in the final set of questions.

A third constraint has to do with examination times. The examinations are administered in only two days, the third and fourth Sunday in the examination month. In my opinion, this is not fair to the students, because they must complete their tests in a very limited time. Students may have to do more than two tests in one day and this may be an reasonable requirement.

The fourth constraint is that analysing items using a statistical approach requires statistical

competencies in the teaching staff, and Universitas Terbuka has only a limited number of experts in this area.

### Item Bank

Developing an item bank is an alternative for assembling tests, because the bank could supply items with their item characteristics to the test developers as soon as tests are needed. Developing new tests without item banking is time consuming and very expensive.

A good item bank is one where the users, in this case the faculty members, can draw items easily. To support this requirement the item bank should have an easy mechanism for retrieving items. The key identifiers in developing tests are the difficulty level, the discriminating power index, and the specification of item objective.

### Recommendations

To develop good items requires student support, good test administration, and expertise in item analysis. Based upon the constraints discussed, before analysing items, I recommend:

a. In relation to the student' support, that students

be motivated to complete the examinations, otherwise the score does not optimize the test information and student ability. For example, before the test time, test supervisors should explain to students that what they do in the examination will absolutely determine their assessment of their competencies. Secondly, the student samples should reach the minimum for the statistical item analysis. The Rasch model needs at least 300 students (Hambleton & Swaminathan, 1985). When the sample does not meet this requirement, for example, the sample is less than 300 students then careful analysis should be considered, because the result of item analysis does not represent to the sample.

- b. Referring to the Examination administration and procedure, Universitas Terbuka should take into account the students' capacity. It is too much to expect students to undertake two or more tests in one day.
- c. The expertise in statistical item analyses is very important in deciding the developing criteria for item selection, based on statistical concepts. So, the statistical item analysis theory and practices

should be provided by item bank staff to the faculty members.

### Item Analysis

In doing item analysis, I recommend the following:

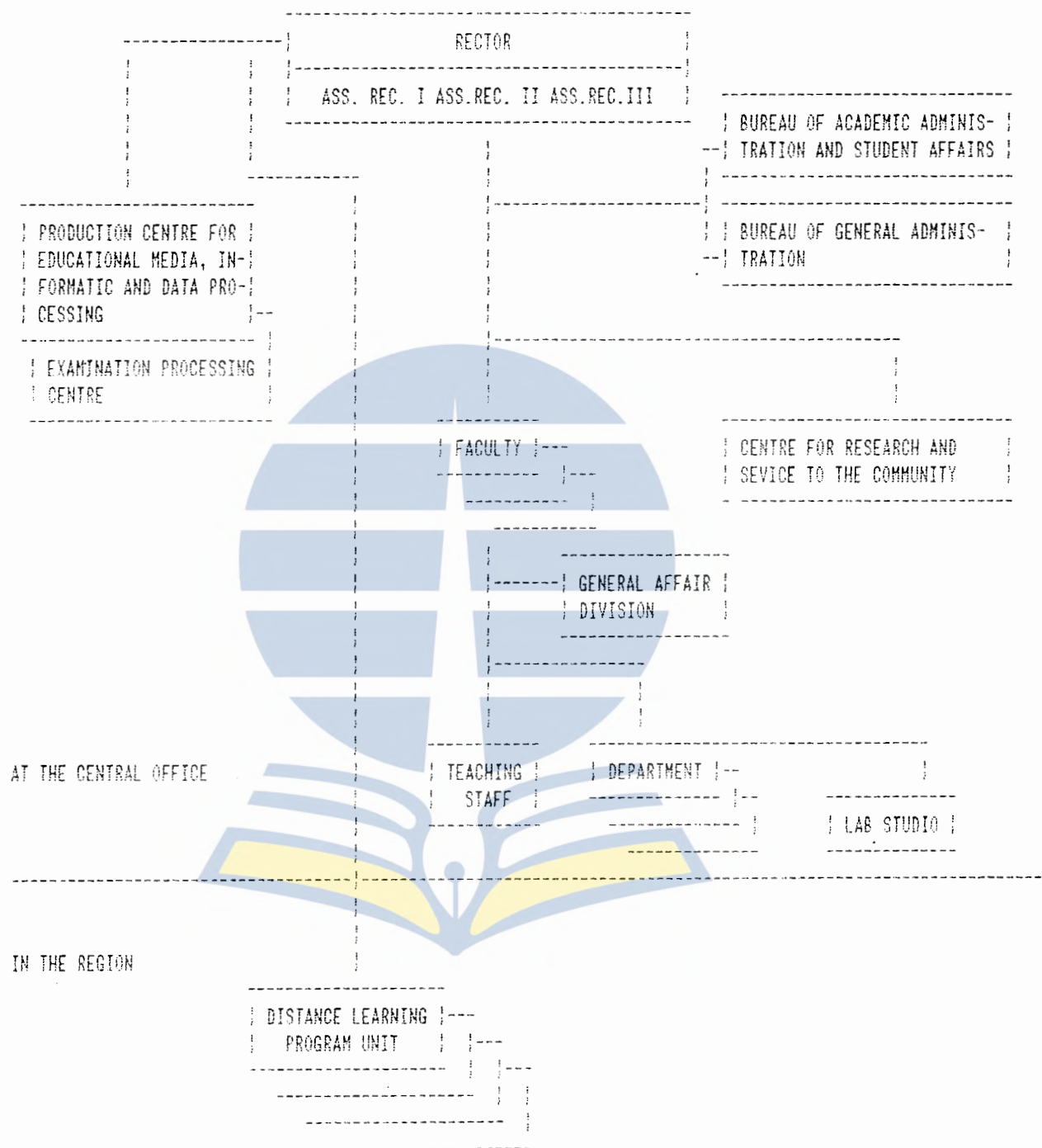
- a. Analyze items using classical item analysis to delete poor items. In classical item analyses information about the p-value and point-biserial correlation are provided also for each option, so distractors analysis can easily be done. The point-biserial index for the correct response is the appropriate indicator for the item discrimination.
- b. To have solid discriminating power, the selected items should be reanalysed using classical item analysis. This revised discrimination index is the one to be kept in the bank.
- c. The Rasch item analysis should be used to calibrate items onto a common difficulty scale. To make the calibration work, at least ten common items should be included in two different tests, because these items will be used to adjust all item difficulties.
- d. The best way to adjust these common items is to average their difficulties. For the rest of items adjust by adding or subtracting half of the mean

difference. On tests in which the common items are more difficult, the remaining items should be adjusted down by half of the mean difference, and the vice versa.



APPENDIX A  
The Structural Organization of  
Universitas Terbuka

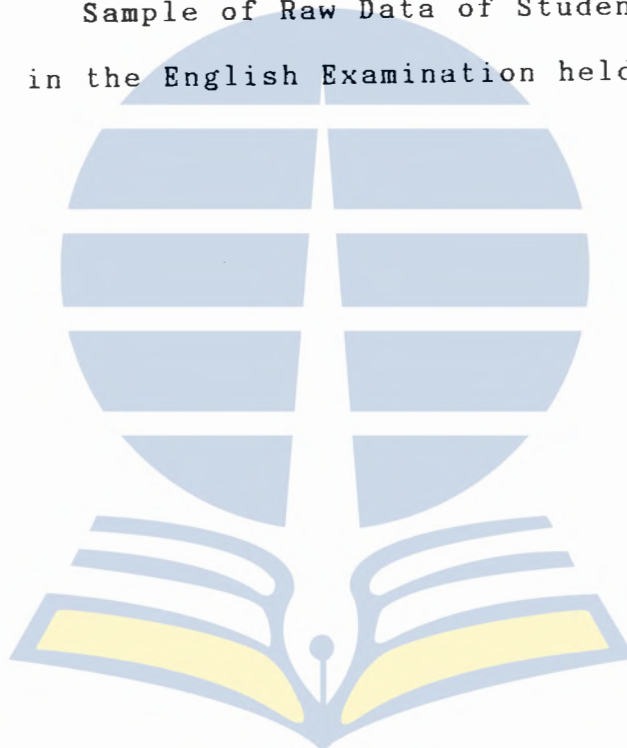






APPENDIX B

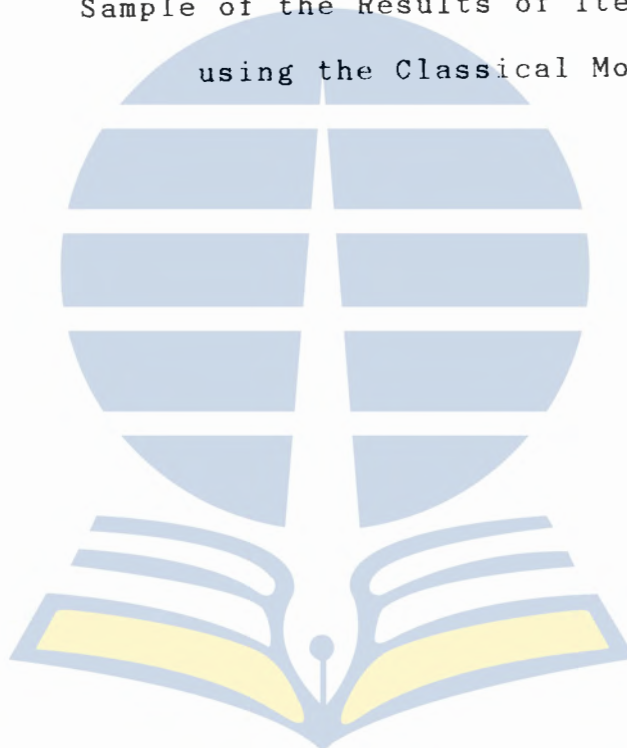
Sample of Raw Data of Student Responses  
in the English Examination held in May 1987



00294786400PING22302904441124243871122C8ADAABCCBDDDBB8A8A8CDAACBAADCC B BDDDBCCABBCCB ACCDABABABADDD
00294368400PING22302708571124243871122C8ADAABCCBDDDBB8A8A8CDAACBAADCC B BDDDBCCABBCCB ACCDABABABADDD
00290622800PING22302805361124240871122C8AABABDCBCDDBAACBAAAABACABADACABACBDDACCABCCBAAAACDCBACACDD68D
00284761500PING22302811601124243871122CAABABDCBDDDBAACBABDCDABABAADCCABBBDDBCCBCCCBAAACCDABABADDDBCD
00283533400PING22300410471124736871122B8A AAAACAACDAB CCB A C B ACAAAC DA CAB CDA CCD CABA A BD
00239347100PING22301212591120201871122C8A BAACABADC BAADBACD DCDD\*ABA\*ADABDBDDBD CBBBACBAAACCAC CAADABC
00234386400PING22301701511144440871122B8AACAABACBDCB8AACBAACCBCBDBBBBAAABADDDCCBACBCBACDADABADAD8A8
00232232300PING22301407571124242871122B8DABAACBACABBBDDCCABACADDD6BABCACBAAACBDDCBBBCBADACCADADCAABDD
00231165300PING22300105531121210871122C8AABAABACBADDDBAABABADABABACDCCABACBDD6CCBACBAAACDCB8ABDDBBB
00231063700PING22301511621121210871122C8BCBAAACBDAADBCACBAAAABBBABABDCCABBBDDBADABCCCBAAAADABACAADD6BB
00228509900PING22302812521124241871122C8AAAACBAAACB8A8A8ADDCDCCDABABCCBAAAADDCB8BBB8BADACBAAACDDDBBC
00227935700PING22301707581121210871122B8ABB8A8ABBB8B88CCDCCBDDABBB8BADAD8AADACB8CBB8A8BDDCDB8ACDD8BD
00227015300PING22302310581114140871122D8CDAACCAAB88B8A8CBA8CDAADAAACDCB8B8DDDCB88C8A8ACADAB8ADDD8DA
00226335500PING22301710511121210871122C8AABAADDADDBAABDBCB8ACAAABADBCA8A8A8DD8CBB8C8A8ACCB8A8AB8D8ACA
00225332200PING22302411501107870871122C8ABB8BC88ADDB8C8B88B8A8A8C8B8ADCC8A8B8DD8C88C8A8ACCB8A8AB8D8ACA
00223872400PING22301703501124241871122C8ACB8B8ACADDD88A8A8A8AADADADCCD8C8AB8D8DD8C88C8B8A8ACCD8AC8B8ADDD8BC
00219274800PING22301008401121210871122C8A A8DCAADD BAACB8A8B8D8ACB8D8ACB8A8B8D8D8C88B8 CAAACACAC8A8A8DD8
00218597100PING22301804591121210871122C8B8B8A8B8CCCB8A8B8ACB8CA8DDB8C8B8A8A8C8B8D8ACB8B8A8ACCC8DA8A8ADD8C8
00216790200PING22302708371124242871122CAAD8A8A8ADDD88A8B8B8A8D8C8B8A8B8D8C8A8A8B8AD8C8C8A8B8C8C8A8C8D8AC8B8DD8B8A
00216452700PING22301908461124242871122C8ACA8B8C8CAD88A8B8A8AD8DC8A8A8A8AC8B8C8DDCC8B8C8B8A8AC8D8C8AC8D8B8B8A
00215799700PING22302510611121210871122B8A8CA
00215062200PING22300808511121210871122C8ABA8B8ACDD8B8C8B8C8B8A8B8AD8C8A8B8B8D8ACD8C8C8B8A8AC8D8A8B8DD8B8DA
00214123600PING22301009581121210871122C8ABB8A8A8B8D8B8A8C8B8A8A8AD8C8B8A8AC8C8A8B8D8D8ACD8C8C8A8AC8A8A8B8D8B8D8
00212488800PING22300000501177770871122B8A8A8A8B8C8D8C8D8C8 B8A8A8B8B8D8A8D8C8C8B8A8A8B8D8D8C8B8A8C8A8A8A8B8C8A8B8C
00211951400PING22301906581121210871122C8ACB8A8ACB8CA8B8A8A8BCA8C8D8C8ACA8C8AD8A8AC8D8C8B8B8B8C8A8AD8C8D8B8B8AD8CA
00210778800PING22300208501121210871122B8ABB8B8B8D8C8DD88A8A8C8\*ACAD8D8C8B8A8C8C8B8C8AD8C8C8B8C8B8A8C8D8A8A8ADDD8B8C
00208969700PING22302712441121210871122C8ADA8B8ACB8ADD8B8C8AD8B8B8B8D8B8A8D8B8A8D8C8B8A8A8D8C8B8C8B8A8C8C8A8C8A8B8AD8AD8
00208801500PING22300101541121210871122CAAD8A8A8AC8AD8A8B8A8A8C8ADDD8C8D8C8A8A8C8D8D8C8C8B8C8B8A8AC8C8A8C8D8A8C8A
00205527800PING22300000581177770871122C8A8A8B8ACB8ADD8B8ACC8B8AC8D8C8A8C8A8A8C8A8B8B8D8AC8C8B8D8C8A8C8D8A8B8C8D8B8D8
00204626100PING22302612561142420871122C8ACB8B8AC8B8D8B8A8B8A8AD8A8C8B8AD8C8C8A8B8D8D8C8C8B8C8B8A8C8D8A8B8C8D8D8D8
00202274600PING22301107411147470871122C8ACB8B8C8B8D8D8C8D8A8AD8A8A8A8AD8C8C8B8B8D8D8AC8C8C8C8B8A8C8D8A8B8ADDD8B8C
00202114200PING22301205521179790871122C8ACB8B8ACB8C8D8B8A8C8B8C8A8A8B8A8A8B8C8B8A8D8D8B8AC8B8C8B8A8C8D8A8B8AC8D8C8B
00198516600PING22300902571144440871122C8A8A8B8ACB8AC8D8DA8C8B8B8B8ADA8A8A8D8C8CA8B8C8D8AD8A8C8C8B8A8AC8D8AC8B8B8B8A
00198488400PING22302803531113410871122C8CADA8ACB8B8B8ACC8CA8AD8B8B8A8B8CA8A8A8B8D8D8C8C8C8B8A8C8A8A8B8C8D8A8B8
00195762600PING22300906481121210871122B8BC8A8D8D8CAD8B8A8C8D8A8A8AD8B8B8C8D8A8B8AD8DD8C8B8C8C8B8A8AC8A8A8C8B8A8D8CA
00194934200PING22303003621124241871122B8BC8A8B8C8C8D8AB8A8A8AC8AD8A8D8B8C8C8D8B8AC8AB8D8C8B8C8B8A8AC8C8B8A8D8C8B8B8A8D8B8B8
00191049300PING22300510461171710871122C8B8DA8A8C8C8B8AD8B8A8C8B8A8A8A8B8D8B8D8C8C8A8B8A8AD8C8A8C8AD8A8A8A8B8D8
00190916600PING22303110351171710871122C8A8B8A8AC8D8D8B8A8C8B8A8A8C8D8AD8B8A8D8C8C8A8C8B8D8ACC8A8C8B8C8A8AC8D8AC8A8B8D8D8
00188288500PING22301405601023230871122B8B8D8B8A8C8D8A8A8C8A8B8A8A8AD8B8A8C8D8C8A8B8B8D8D8C8B8C8C8A8A8A8C8D8A8C8D8D8
00186441100PING22300801601144440871122C8A8A8A8A8AC8D8AD8B8A8C8A8A8AD8B8B8DA8CC8A8A8B8D8D8C8C8A8B8D8B8A8AD8A8A8A8C8D8B8
00186414200PING22301809611144440871122B8A8A8A8A8B8D8B8A8A8A8A8AC8ADA8C8B8AD8C8D8B8D8D8ACC8C8B8C8A8AC8DA8A8A8AD8C8D8
00179338400PING22301007561144440871122CA8B8A8A8AC8D8D8B8A8C8A8A8AD8B8D8AC8D8C8C8A8B8D8C8B8C8A8AC8D8AC8A8AD8D8B8A8
00179061900PING22302303491142423871122B8A8B8A8A8AD8A8A8C8A8A8AD8B8A8A8C8A8A8D8C8B8C8B8B8C8B8A8A8C8E8B8C8AD8B8B8
00177197300PING22300703601142420871122C8B8D8A8B8AC8C8D8B8AC8B8B8AC8D8A8B8A8B8B8C8A8A8D8D8AC8B8A8C8A8A8A8AD8A8D8A8
00171065500PING22302712511182821871122C8DA8D8A8B8C8D8B8B8D8A8A8C8D8D8AC8D8A8D8B8AD8C8B8C8C8A8D8C8A8A8C8D8A8B8D8

## APPENDIX C

Sample of the Results of Item Analyses  
using the Classical Model



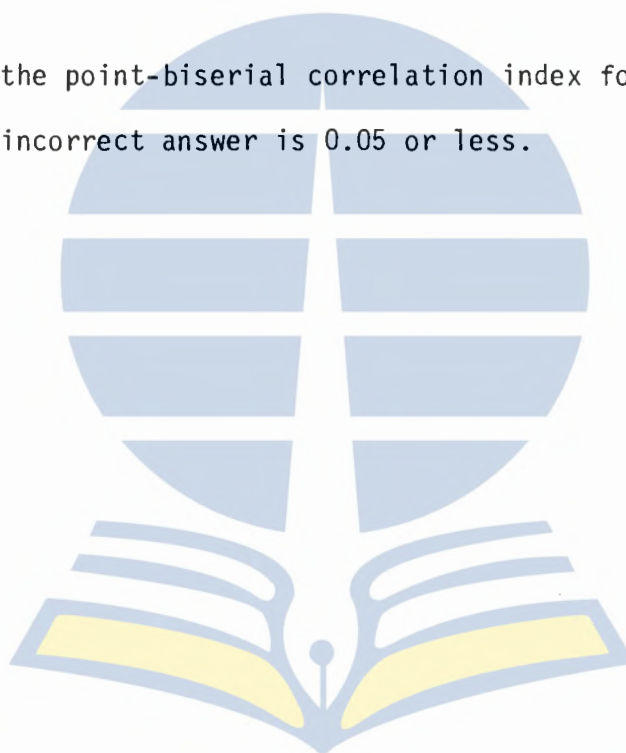
Seq. No.	Scale -Item	Item Statistics			Alternative Statistics				Key
		Prop. Correct	Biser.	Point Biser.	Alt.	Prop. Endorsing	Biser.	Point Biser.	
1	0-1	0.609	0.376	0.296	A	0.130	0.067	0.042	✓
					B	0.152	-0.181	-0.119	
					C	0.609	0.376	0.296	* OK
					D	0.087	-0.338	-0.190	
					Other	0.022	-1.000	-0.428	
2	0-2	0.130	-0.175	-0.110	A	0.348	-0.093	-0.072	
					B	0.130	-0.175	-0.110	*
					C	0.196	-0.122	-0.085	
					D	0.283	0.539	0.405	? X
					Other	0.043	-0.837	-0.378	-
CHECK THE KEY B was specified, D works better									
3	0-3	0.304	0.243	0.185	A	0.348	0.016	0.013	
					B	0.304	0.243	0.185	* X
					C	0.152	0.083	0.055	
					D	0.152	-0.197	-0.129	
					Other	0.043	-0.698	-0.315	
4	0-4	0.565	0.111	0.088	A	0.174	-0.110	-0.074	
					B	0.565	0.111	0.088	*
					C	0.217	0.222	0.158	? X
					D	0.000	-9.000	-9.000	-
					Other	0.043	-0.877	-0.396	
CHECK THE KEY B was specified, C works better									
5	0-5	0.239	0.003	0.002	A	0.196	0.475	0.331	? X
					B	0.239	0.003	0.002	*
					C	0.109	0.008	0.005	
					D	0.391	-0.176	-0.138	
					Other	0.065	-0.521	-0.268	
CHECK THE KEY B was specified, A works better									
6	0-6	0.283	0.962	0.722	A	0.391	-0.281	-0.221	
					B	0.109	-0.327	-0.196	
					C	0.174	-0.274	-0.185	
					D	0.283	0.962	0.722	* ✓
					Other	0.043	-0.936	-0.423	
CHECK THE KEY B was specified, D works better									OK

Appendix D  
Results of Item Analysis of the Reduced Item  
Collection using the Classical Model



NOTE: The items analyzed in this Appendix are those that met the following criteria in the original analysis:

- 1) the point-biserial correlation index for the correct answer is 0.20 or higher;
- 2) all incorrect options have been chosen by some students;
- 3) the point-biserial correlation index for each incorrect answer is 0.05 or less.



## 1. ENGLISH IN DECEMBER 1986 EXAMINATION

Seq. No.	Scale -Item	Item Statistics			Alternative Statistics				
		Prop. Correct	Biser.	Point Biser.	Alt.	Prop. Endorsing	Biser.	Point Biser.	Key
1	0-1	0.622	0.357	0.280	A	0.133	0.024	0.015	
					B	0.156	-0.342	-0.226	
					C	0.622	0.357	0.280	*
					D	0.089	-0.368	-0.208	
					Other	0.000	-9.000	-9.000	
6	0-2	0.289	1.000	0.778	A	0.400	-0.394	-0.311	
					B	0.111	-0.311	-0.187	
					C	0.178	-0.455	-0.310	
					D	0.289	1.000	0.778	*
					Other	0.022	-0.438	-0.157	
11	0-3	0.289	0.267	0.201	A	0.578	0.003	0.002	
					B	0.022	-0.438	-0.157	
					C	0.111	-0.364	-0.219	
					D	0.289	0.267	0.201	*
					Other	0.000	-9.000	-9.000	
13	0-4	0.511	0.690	0.550	A	0.133	-0.397	-0.251	
					B	0.156	-0.384	-0.253	
					C	0.511	0.690	0.550	*
					D	0.200	-0.350	-0.245	
					Other	0.000	-9.000	-9.000	
16	0-5	0.622	0.666	0.522	A	0.622	0.666	0.522	*
					B	0.178	-0.300	-0.204	
					C	0.156	-0.539	-0.355	
					D	0.044	-0.494	-0.225	
					Other	0.000	-9.000	-9.000	
17	0-6	0.489	0.437	0.349	A	0.489	0.437	0.349	*
					B	0.333	-0.394	-0.304	
					C	0.067	-0.045	-0.023	
					D	0.067	-0.486	-0.252	
					Other	0.044	0.399	0.182	

(continued)

Seq. No.	Scale -Item	Item Statistics			Alternative Statistics			
		Prop. Correct	Biser.	Point Biser.	Alt.	Prop. Endorsing	Biser.	Point Biser. Key
18	0-7	0.622	0.675	0.529	A	0.178	-0.622	-0.424
					B	0.622	0.675	*
					C	0.067	-0.071	-0.037
					D	0.133	-0.397	-0.251
					Other	0.000	-9.000	-9.000
19	0-8	0.422	0.820	0.650	A	0.089	-0.118	-0.067
					B	0.244	-0.609	-0.445
					C	0.200	-0.398	-0.278
					D	0.422	0.820	0.650 *
					Other	0.044	0.006	0.003
26	0-9	0.311	0.638	0.487	A	0.222	-0.339	-0.243
					B	0.311	0.638	0.487 *
					C	0.133	-0.054	-0.034
					D	0.267	-0.387	-0.288
					Other	0.067	0.111	0.057
27	0-10	0.600	0.516	0.407	A	0.111	-0.453	-0.273
					B	0.222	-0.215	-0.154
					C	0.600	0.516	0.407 *
					D	0.067	-0.382	-0.198
					Other	0.000	-9.000	-9.000
32	0-11	0.578	0.577	0.457	A	0.200	-0.254	-0.178
					B	0.022	-0.501	-0.188
					C	0.578	0.577	0.457 *
					D	0.178	-0.519	-0.353
					Other	0.022	0.132	0.048
33	0-12	0.711	0.607	0.458	A	0.133	-0.443	-0.281
					B	0.711	0.607	0.458 *
					C	0.067	-0.538	-0.279
					D	0.089	-0.264	-0.149
					Other	0.000	-9.000	-9.000



(continued)

Seq. No.	Scale -Item	Item Statistics			Alternative Statistics				
		Prop. Correct	Biser.	Point Biser.	Alt.	Prop. Endorsing	Biser.	Point Biser.	Key
35	0-13	0.467	0.488	0.389	A	0.200	-0.482	-0.337	
					B	0.467	0.488	0.389	*
					C	0.289	-0.087	-0.065	
					D	0.044	-0.315	-0.144	
					Other	0.000	-9.000	-9.000	
37	0-14	0.333	0.538	0.415	A	0.222	-0.125	-0.090	
					B	0.333	0.538	0.415	*
					C	0.356	-0.158	-0.123	
					D	0.089	-0.618	-0.349	
					Other	0.000	-9.000	-9.000	
38	0-15	0.889	0.346	0.209	A	0.889	0.346	0.209	*
					B	0.022	-0.058	-0.021	
					C	0.022	-0.375	-0.134	
					D	0.067	-0.330	-0.171	
					Other	0.000	-9.000	-9.000	
42	0-16	0.356	0.715	0.556	A	0.333	-0.550	-0.424	
					B	0.067	-0.149	-0.077	
					C	0.356	0.715	0.556	*
					D	0.244	-0.150	-0.109	
					Other	0.000	-9.000	-9.000	
44	0-17	0.667	0.698	0.538	A	0.667	0.698	0.538	*
					B	0.156	-0.427	-0.281	
					C	0.067	-0.564	-0.292	
					D	0.111	-0.417	-0.251	
					Other	0.000	-9.000	-9.000	
49	0-18	0.689	0.776	0.592	A	0.689	0.776	0.592	*
					B	0.111	-0.488	-0.294	
					C	0.067	-0.408	-0.212	
					D	0.133	-0.599	-0.380	
					Other	0.000	-9.000	-9.000	

(continued)

Seq. No.	Scale -Item	Item Statistics			Alternative Statistics				
		Prop. Correct	Biser.	Point Biser.	Alt.	Prop. Endorsing	Biser.	Point Biser.	Key
50	0-19	0.511	0.539	0.430	A	0.511	0.539	0.430	*
					B	0.156	-0.286	-0.189	
					C	0.267	-0.377	-0.290	
					D	0.067	-0.175	-0.091	
					Other	0.000	-9.000	-9.000	
52	0-20	0.467	0.834	0.665	A	0.267	-0.367	-0.273	
					B	0.044	-0.494	-0.225	
					C	0.467	0.834	0.665	*
					D	0.200	-0.529	-0.371	
					Other	0.022	-0.311	-0.112	
58	0-21	0.489	0.572	0.456	A	0.178	-0.429	-0.292	
					B	0.244	-0.278	-0.203	
					C	0.489	0.572	0.456	*
					D	0.022	-0.501	-0.180	
					Other	0.067	-0.019	-0.010	
59	0-22	0.578	0.603	0.477	A	0.022	-0.184	-0.066	
					B	0.578	0.603	0.477	*
					C	0.044	-0.423	-0.192	
					D	0.267	-0.581	-0.432	
					Other	0.089	0.028	0.016	
60	0-23	0.422	0.357	0.283	A	0.267	-0.203	-0.151	
					B	0.200	-0.038	-0.027	
					C	0.067	-0.512	-0.265	
					D	0.422	0.357	0.283	*
					Other	0.044	0.042	0.019	
61	0-24	0.178	0.898	0.612	A	0.222	-0.395	-0.283	
					B	0.067	-0.408	-0.212	
					C	0.489	-0.169	-0.134	
					D	0.178	0.898	0.612	*
					Other	0.044	0.042	0.019	

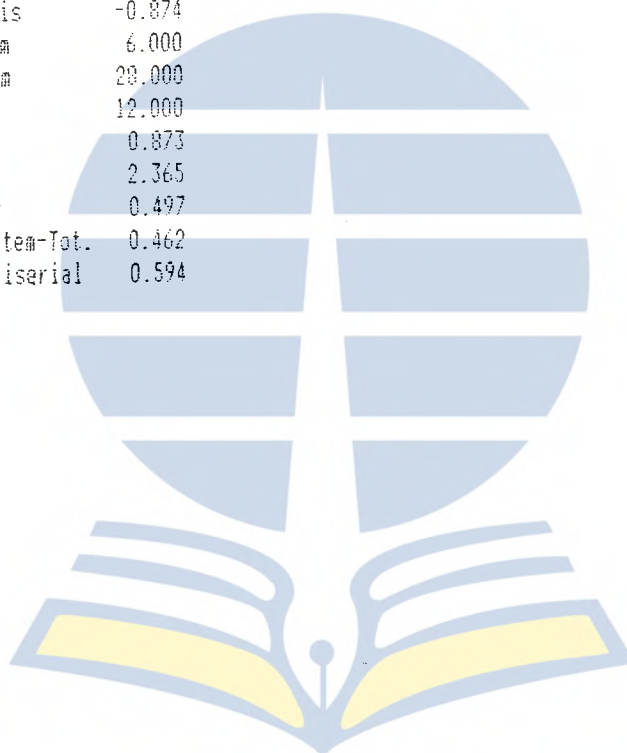
(continued)

Seq. No.	Scale -Item	Item Statistics			Alternative Statistics				
		Prop. Correct	Biser.	Point Biser.	Alt.	Prop. Endorsing	Biser.	Point Biser.	Key
62	0-25	0.511	0.656	0.524	A	0.511	0.656	0.524	*
					B	0.289	-0.430	-0.324	
					C	0.089	-0.451	-0.255	
					D	0.067	-0.356	-0.185	
					Other	0.044	0.042	0.019	
63	0-26	0.444	0.584	0.464	A	0.356	-0.275	-0.214	
					B	0.444	0.584	0.464	*
					C	0.133	-0.521	-0.330	
					D	0.022	-0.375	-0.134	
					Other	0.044	0.042	0.019	
65	0-27	0.489	0.201	0.161	A	0.489	0.201	0.161	*
					B	0.244	-0.118	-0.086	
					C	0.178	-0.120	-0.081	
					D	0.022	-0.058	-0.021	
					Other	0.067	-0.071	-0.037	
69	0-28	0.467	0.446	0.355	A	0.467	0.446	0.355	*
					B	0.156	-0.272	-0.179	
					C	0.178	-0.261	-0.178	
					D	0.067	-0.227	-0.117	
					Other	0.133	-0.070	-0.044	
70	0-29	0.467	0.606	0.483	A	0.467	0.606	0.483	*
					B	0.067	-0.278	-0.144	
					C	0.200	-0.242	-0.169	
					D	0.133	-0.568	-0.360	
					Other	0.133	-0.070	-0.044	
78	0-30	0.422	0.726	0.575	A	0.422	0.726	0.575	*
					B	0.156	-0.427	-0.281	
					C	0.067	-0.304	-0.158	
					D	0.244	-0.460	-0.336	
					Other	0.111	0.008	0.005	

(continued)

## Scale Statistics

Scale:	0
N of Items	30
N of Examinees	45
Mean	14.911
Variance	43.903
Std. Dev.	6.626
Skew	0.613
Kurtosis	-0.874
Minimum	6.000
Maximum	28.000
Median	12.000
Alpha	0.873
SEM	2.365
Mean P	0.497
Mean Item-Tot.	0.462
Mean Biserial	0.594



## 2. ENGLISH IN MAY 1987 EXAMINATION

Seq. No.	Scale -Item	Item Statistics			Alternative Statistics				
		Prop. Correct	Biser.	Point Biser.	Alt.	Prop. Endorsing	Biser.	Point Biser.	Key
2	0-1	0.808	0.536	0.372	A	0.151	-0.295	-0.193	
					B	0.808	0.536	0.372	*
					C	0.027	-0.848	-0.328	
					D	0.014	-0.682	-0.205	
					Other	0.000	-9.000	-9.000	
3	0-2	0.658	0.684	0.529	A	0.658	0.684	0.529	*
					B	0.205	-0.356	-0.251	
					C	0.055	-0.305	-0.149	
					D	0.082	-0.765	-0.422	
					Other	0.000	-9.000	-9.000	
6	0-3	0.849	0.958	0.626	A	0.849	0.958	0.626	*
					B	0.014	-0.805	-0.242	
					C	0.041	-0.536	-0.238	
					D	0.082	-0.736	-0.407	
					Other	0.014	-1.000	-0.317	
7	0-4	0.534	0.776	0.618	A	0.329	-0.355	-0.273	
					B	0.534	0.776	0.618	*
					C	0.014	-0.559	-0.168	
					D	0.110	-0.659	-0.396	
					Other	0.014	-1.000	-0.317	
8	0-5	0.575	0.472	0.374	A	0.575	0.472	0.374	*
					B	0.082	-0.613	-0.339	
					C	0.219	-0.074	-0.053	
					D	0.110	-0.177	-0.106	
					Other	0.014	-1.000	-0.317	
9	0-6	0.712	0.444	0.334	A	0.151	-0.307	-0.201	
					B	0.068	-0.285	-0.149	
					C	0.712	0.444	0.334	*
					D	0.055	-0.046	-0.023	
					Other	0.014	-1.000	-0.317	

(continued)

Seq. No.	Scale -Item	Item Statistics			Alternative Statistics				Key
		Prop. Correct	Biser.	Point Biser.	Alt.	Prop. Endorsing	Biser.	Point Biser.	
10	0-7	0.644	0.218	0.170	A	0.233	0.028	0.021	
					B	0.644	0.218	0.170	*
					C	0.055	-0.331	-0.161	
					D	0.055	-0.150	-0.073	
					Other	0.014	-1.000	-0.317	
11	0-8	0.301	0.609	0.462	A	0.288	-0.208	-0.157	
					B	0.082	-0.207	-0.114	
					C	0.315	-0.205	-0.157	
					D	0.301	0.609	0.462	*
					Other	0.014	-1.000	-0.317	
13	0-9	0.630	0.678	0.530	A	0.137	-0.369	-0.236	
					B	0.110	-0.353	-0.212	
					C	0.110	-0.384	-0.230	
					D	0.630	0.678	0.530	*
					Other	0.014	-1.000	-0.317	
15	0-10	0.877	0.808	0.501	A	0.014	-0.805	-0.242	
					B	0.877	0.808	0.501	*
					C	0.027	-0.757	-0.293	
					D	0.068	-0.394	-0.206	
					Other	0.014	-1.000	-0.317	
16	0-11	0.562	0.559	0.444	A	0.562	0.559	0.444	*
					B	0.192	-0.526	-0.364	
					C	0.205	-0.074	-0.052	
					D	0.014	-0.313	-0.094	
					Other	0.027	-0.711	-0.275	
17	0-12	0.699	0.680	0.516	A	0.699	0.680	0.516	*
					B	0.041	-0.716	-0.318	
					C	0.219	-0.356	-0.254	
					D	0.027	-0.507	-0.196	
					Other	0.014	-1.000	-0.317	

(continued)

Seq. No.	Scale	Item	Item Statistics					Alternative Statistics				
			Point	Prop.	Correct	Biserr. Prop.	Biserr. Point	Point	Prop.	Endorsing	Biserr. Point	Biserr. Key
19	0-13		0.607	0.780	0.644	0.780	0.260	0.644	0.780	0.607	*	
		A					-0.470	0.780	-0.347			
		B					0.644	0.780	0.607			
		C					0.027	-0.575	-0.222			
		D					0.055	-0.590	-0.287			
		Other					0.014	-1.000	-0.317			
24	0-14		0.521	0.667	0.630	0.667	0.096	-0.164	-0.095			
		A					-0.096	-0.164	-0.095			
		B					0.137	-0.533	-0.340			
		C					0.110	-0.330	-0.198			
		D					0.630	0.667	0.521		*	
		Other					0.027	-0.711	-0.275			
30	0-15		0.291	0.365	0.466	0.365	0.466	0.365	0.291		*	
		A					0.466	0.365	0.291			
		B					0.247	-0.130	-0.095			
		C					0.123	-0.111	-0.069			
		D					0.137	-0.166	-0.106			
		Other					0.027	-0.711	-0.275			
31	0-16		0.400	0.503	0.452	0.503	0.301	-0.067	-0.051		*	
		A					0.301	-0.067	-0.051			
		B					0.178	-0.294	-0.200			
		C					0.041	-0.667	-0.296			
		D					0.452	0.503	0.400		*	
		Other					0.027	-0.643	-0.249			
32	0-17		0.514	0.648	0.575	0.648	0.137	0.043	0.027		*	
		A					0.137	0.043	0.027			
		B					0.123	-0.379	-0.235			
		C					0.575	0.648	0.514		*	
		D					0.110	-0.659	-0.396			
		Other					0.055	-0.564	-0.275			
33	0-18		0.614	0.773	0.438	0.773	0.247	-0.271	-0.198		*	
		A					0.247	-0.271	-0.198			
		B					0.192	-0.273	-0.190			
		C					0.438	0.773	0.614		*	
		D					0.110	-0.575	-0.345			
		Other					0.014	-1.000	-0.317			

Seq. No.	Scale	Item No.	Item Statistics				Alternative Statistics			
			Point	Prop.	Correct Biser.	Prop. Biser.	Point	Prop.	Alt. Endorsing Biser.	Point Biser. Key
35	0-19		0.712	0.570	0.430	0.164	-0.155	-0.103		
		A				0.712	0.570	0.430		
		B				0.212	0.570	0.430		*
		C				0.055	-0.654	-0.319		
		D				0.055	-0.421	-0.205		
		Other				0.014	-1.000	-0.317		
36	0-20		0.616	0.440	0.345	0.616	0.440	0.345		*
		A				0.616	0.440	0.345		
		B				0.219	-0.099	-0.070		
		C				0.068	-0.513	-0.268		
		D				0.068	-0.361	-0.189		
		Other				0.027	-0.370	-0.143		
38	0-21		0.630	0.503	0.393	0.123	-0.104	-0.065		*
		A				0.123	-0.104	-0.065		
		B				0.630	0.503	0.393		
		C				0.164	-0.300	-0.200		
		D				0.055	-0.447	-0.218		
		Other				0.027	-0.711	-0.275		
39	0-22		0.658	0.649	0.502	0.137	-0.160	-0.102		*
		A				0.137	-0.160	-0.102		
		B				0.082	-0.689	-0.381		
		C				0.096	-0.316	-0.183		
		D				0.658	0.649	0.502		
		Other				0.027	-0.711	-0.275		
40	0-23		0.644	0.415	0.323	0.055	-0.189	-0.092		*
		A				0.055	-0.189	-0.092		
		B				0.205	-0.029	-0.020		
		C				0.082	-0.585	-0.323		
		D				0.644	0.415	0.323		
		Other				0.014	-1.000	-0.317		
42	0-24		0.562	0.603	0.479	0.151	-0.258	-0.169		*
		A				0.151	-0.258	-0.169		
		B				0.123	-0.259	-0.161		
		C				0.562	0.603	0.479		
		D				0.123	-0.357	-0.222		
		Other				0.041	-0.585	-0.260		



(continued)

Seq. No.	Scale -Item	Item Statistics			Alternative Statistics				
		Prop. Correct	Biser. Biser.	Point Biser.	Alt.	Prop. Endorsing	Biser. Biser.	Point Biser.	Key
43	0-25	0.822	0.723	0.493	A	0.041	-0.764	-0.339	
					B	0.055	-0.654	-0.319	
					C	0.822	0.723	0.493	*
					D	0.068	-0.089	-0.047	
					Other	0.014	-1.000	-0.317	
46	0-26	0.452	0.466	0.371	A	0.068	-0.796	-0.416	
					B	0.425	0.022	0.018	
					C	0.452	0.466	0.371	*
					D	0.041	-0.585	-0.260	
					Other	0.014	-1.000	-0.317	
48	0-27	0.699	0.804	0.610	A	0.041	-0.520	-0.231	
					B	0.151	-0.528	-0.345	
					C	0.699	0.804	0.610	*
					D	0.082	-0.500	-0.276	
					Other	0.027	-0.552	-0.213	
49	0-28	0.685	0.557	0.426	A	0.151	-0.185	-0.121	
					B	0.685	0.557	0.426	*
					C	0.110	-0.292	-0.175	
					D	0.041	-0.716	-0.318	
					Other	0.014	-1.000	-0.317	
50	0-29	0.767	0.765	0.553	A	0.767	0.765	0.553	*
					B	0.068	-0.643	-0.336	
					C	0.082	-0.254	-0.140	
					D	0.055	-0.784	-0.382	
					Other	0.027	-0.370	-0.143	
51	0-30	0.685	0.767	0.586	A	0.685	0.767	0.586	*
					B	0.137	-0.428	-0.273	
					C	0.082	-0.471	-0.260	
					D	0.068	-0.470	-0.245	
					Other	0.027	-0.711	-0.275	

(continued)

Seq. No.	Scale -Item	Item Statistics			Alternative Statistics				
		Prop. Correct	Biser.	Point Biser.	Alt.	Prop. Endorsing	Biser.	Point Biser.	Key
53	0-31	0.534	0.667	0.532	A	0.370	-0.283	-0.221	
					B	0.055	-0.758	-0.369	
					C	0.534	0.667	0.532	*
					D	0.027	-0.598	-0.231	
					Other	0.014	-1.000	-0.317	
54	0-32	0.452	0.666	0.530	A	0.027	-0.620	-0.240	
					B	0.096	-0.561	-0.324	
					C	0.411	-0.237	-0.188	
					D	0.452	0.666	0.530	*
					Other	0.014	-1.000	-0.317	
55	0-33	0.630	0.655	0.512	A	0.630	0.655	0.512	*
					B	0.151	-0.295	-0.193	
					C	0.151	-0.442	-0.289	
					D	0.027	-0.416	-0.161	
					Other	0.041	-0.553	-0.245	
57	0-34	0.658	0.797	0.617	A	0.658	0.797	0.617	*
					B	0.164	-0.329	-0.220	
					C	0.041	-0.813	-0.361	
					D	0.082	-0.566	-0.313	
					Other	0.055	-0.486	-0.237	
58	0-35	0.616	0.707	0.555	A	0.192	-0.689	-0.477	
					B	0.616	0.707	0.555	*
					C	0.123	0.051	0.031	
					D	0.041	-0.585	-0.260	
					Other	0.027	-0.643	-0.249	
61	0-36	0.562	0.541	0.430	A	0.315	-0.064	-0.049	
					B	0.068	-0.676	-0.353	
					C	0.041	-0.732	-0.325	
					D	0.562	0.541	0.430	*
					Other	0.014	-1.000	-0.317	

(continued)

Seq. No.	Scale -Item	Item Statistics			Alternative Statistics				
		Prop. Correct	Biser. Biser.	Point Biser.	Alt.	Prop. Endorsing	Biser. Biser.	Point Biser.	Key
62	0-37	0.589	0.809	0.640	A	0.274	-0.486	-0.363	
					B	0.096	-0.578	-0.334	
					C	0.014	-0.272	-0.082	
					D	0.589	0.809	0.640	*
					Other	0.027	-0.711	-0.275	
67	0-38	0.274	0.517	0.386	A	0.274	0.517	0.386	†
					B	0.260	-0.302	-0.223	
					C	0.096	-0.341	-0.197	
					D	0.356	0.055	0.043	
					Other	0.014	-1.000	-0.317	
69	0-39	0.534	0.458	0.365	A	0.534	0.458	0.365	*
					B	0.110	-0.307	-0.184	
					C	0.151	-0.105	-0.069	
					D	0.137	-0.356	-0.227	
					Other	0.068	-0.165	-0.086	
73	0-40	0.534	0.678	0.540	A	0.534	0.678	0.540	*
					B	0.137	-0.566	-0.361	
					C	0.151	-0.142	-0.093	
					D	0.123	-0.407	-0.252	
					Other	0.055	-0.266	-0.130	
76	0-41	0.438	0.606	0.481	A	0.164	-0.033	-0.022	
					B	0.192	-0.373	-0.259	
					C	0.438	0.606	0.481	*
					D	0.110	-0.430	-0.258	
					Other	0.096	-0.282	-0.163	
77	0-42	0.411	0.449	0.355	A	0.151	-0.160	-0.105	
					B	0.137	-0.180	-0.111	
					C	0.411	0.449	0.355	†
					D	0.219	-0.196	-0.140	
					Other	0.082	-0.263	-0.146	

(continued)

Seq. No.	Scale -Item	Item Statistics			Alternative Statistics			
		Prop. Correct	Biser. Biser.	Point Biser.	Alt.	Prop. Endorsing	Biser. Biser.	Point Biser. Key
78	0-43	0.425	0.564	0.447	A	0.096	-0.358	-0.207
					B	0.137	-0.337	-0.215
					C	0.233	-0.098	-0.071
					D	0.425	0.564	0.447 *
					Other	0.110	-0.300	-0.180

## Scale Statistics

Scale:	0
N of Items	43
N of Examinees	73
Mean	25.644
Variance	91.161
Std. Dev.	9.548
Skew	-0.527
Kurtosis	-0.484
Minimum	0.000
Maximum	41.000
Median	27.000
Alpha	0.916
SEM	2.766
Mean P	0.596
Mean Item-Tot.	0.473
Mean Biserial	0.617

## 3. MATHEMATICS IN DECEMBER 1986 EXAMINATION

Seq. No.	Scale -Item	Item Statistics			Alternative Statistics				
		Prop. Correct	Biser.	Point Biser.	Alt.	Prop. Endorsing	Biser.	Point Biser.	Key
1	0-1	0.440	0.548	0.436	A	0.096	-0.626	-0.362	
					B	0.380	-0.168	-0.131	
					C	0.440	0.548	0.436	*
					D	0.078	-0.248	-0.135	
					Other	0.006	-0.544	-0.120	
2	0-2	0.542	0.496	0.395	A	0.542	0.496	0.395	*
					B	0.151	-0.250	-0.163	
					C	0.205	-0.302	-0.212	
					D	0.102	-0.295	-0.174	
					Other	0.000	-9.000	-9.000	
3	0-3	0.422	0.465	0.368	A	0.301	-0.107	-0.081	
					B	0.422	0.465	0.368	*
					C	0.199	-0.414	-0.289	
					D	0.060	-0.163	-0.082	
					Other	0.018	-0.214	-0.071	
5	0-4	0.687	0.596	0.455	A	0.687	0.596	0.455	*
					B	0.108	-0.398	-0.238	
					C	0.169	-0.410	-0.276	
					D	0.024	-0.396	-0.146	
					Other	0.012	-0.358	-0.103	
6	0-5	0.916	0.571	0.318	A	0.054	-0.582	-0.282	
					B	0.006	-0.830	-0.183	
					C	0.012	-0.280	-0.080	
					D	0.916	0.571	0.318	*
					Other	0.012	-0.046	-0.013	
7	0-6	0.759	0.524	0.382	A	0.114	-0.336	-0.204	
					B	0.759	0.524	0.382	*
					C	0.078	-0.323	-0.176	
					D	0.030	-0.733	-0.293	
					Other	0.018	-0.021	-0.007	

(continued)

Seq. No.	Scale -Item	Item Statistics			Alternative Statistics				
		Prop. Correct	Biser.	Point Biser.	Alt.	Prop. Endorsing	Biser.	Point Biser.	Key
8	0-7	0.675	0.336	0.258	A	0.223	-0.221	-0.158	
					B	0.084	-0.263	-0.146	
					C	0.675	0.336	0.258	*
					D	0.012	-0.163	-0.047	
					Other	0.006	-0.544	-0.120	
9	0-8	0.596	0.321	0.253	A	0.187	-0.202	-0.139	
					B	0.084	-0.200	-0.111	
					C	0.596	0.321	0.253	*
					D	0.127	-0.144	-0.090	
					Other	0.006	-0.544	-0.120	
11	0-9	0.325	0.455	0.350	A	0.325	0.455	0.350	*
					B	0.102	-0.226	-0.133	
					C	0.133	-0.462	-0.292	
					D	0.416	-0.068	-0.054	
					Other	0.024	0.035	0.013	
12	0-10	0.807	0.799	0.554	A	0.060	-0.521	-0.262	
					B	0.030	-0.805	-0.322	
					C	0.807	0.799	0.554	*
					D	0.090	-0.532	-0.302	
					Other	0.012	-0.475	-0.136	
14	0-11	0.470	0.482	0.384	A	0.217	-0.076	-0.054	
					B	0.470	0.482	0.384	*
					C	0.102	-0.247	-0.145	
					D	0.205	-0.409	-0.288	
					Other	0.006	-0.544	-0.120	
15	0-12	0.693	0.592	0.451	A	0.693	0.592	0.451	*
					B	0.084	-0.563	-0.314	
					C	0.108	-0.300	-0.179	
					D	0.102	-0.308	-0.182	
					Other	0.012	-0.319	-0.092	

(continued)

Seq. No.	Scale -Item	Correct Prop.	Item Statistics				Alternative Statistics				
			Point	Biserial	Biserial	Point	Alt. Endorsing Prop.	Biserial	Biserial	Point	
17	0-13	0.301	0.412	0.313	0.289	-0.187	-0.141	0.301	0.412	0.313	*
								A	0.163	-0.005	-0.003
								C	0.169	-0.328	-0.220
								D	0.078	0.028	0.015
								Other			
19	0-14	0.476	0.551	0.439	0.476	0.551	0.439	A	0.367	-0.519	-0.406
								B	0.054	-0.148	-0.072
								C	0.096	-0.039	-0.023
								D	0.006	-0.042	-0.009
								Other			
20	0-15	0.373	0.521	0.408	0.373	0.521	0.408	A	0.373	0.521	0.408
								B	0.427	-0.341	-0.408
								C	0.048	-0.070	-0.033
								D	0.000	-9.000	-9.000
								Other			
21	0-16	0.789	0.583	0.413	0.789	0.583	0.413	A	0.024	-0.655	-0.242
								B	0.042	-0.505	-0.226
								C	0.789	0.583	0.413
								D	0.145	-0.377	-0.244
								Other	0.000	-9.000	-9.000
23	0-17	0.807	0.625	0.434	0.807	0.625	0.434	A	0.114	-0.501	-0.304
								B	0.807	0.625	0.434
								C	0.036	-0.486	-0.207
								D	0.042	-0.396	-0.177
								Other	0.000	-9.000	-9.000
25	0-18	0.735	0.649	0.482	0.735	0.649	0.482	A	0.042	-0.192	-0.086
								B	0.030	-0.877	-0.350
								C	0.193	-0.495	-0.344
								D	0.735	0.649	0.482
								Other	0.000	-9.000	-9.000

(continued)

Seq. No.	Scale -Item	Item Statistics			Alternative Statistics				
		Prop. Correct	Biser.	Point Biser.	Alt.	Prop. Endorsing	Biser.	Point Biser.	Key
26	0-19	0.584	0.736	0.582	A	0.584	0.736	0.582	*
					B	0.151	-0.401	-0.262	
					C	0.211	-0.422	-0.299	
					D	0.048	-0.729	-0.341	
					Other	0.006	0.101	0.022	
27	0-20	0.560	0.654	0.520	A	0.211	-0.409	-0.290	
					B	0.560	0.654	0.520	*
					C	0.090	-0.479	-0.272	
					D	0.120	-0.310	-0.191	
					Other	0.018	0.006	0.002	
28	0-21	0.572	0.582	0.462	A	0.169	-0.245	-0.165	
					B	0.139	-0.396	-0.253	
					C	0.572	0.582	0.462	*
					D	0.096	-0.468	-0.271	
					Other	0.024	0.014	0.005	
30	0-22	0.289	0.310	0.234	A	0.277	-0.201	-0.151	
					B	0.289	0.310	0.234	*
					C	0.277	-0.044	-0.033	
					D	0.108	-0.201	-0.120	
					Other	0.048	0.138	0.064	
36	0-23	0.253	0.410	0.301	A	0.253	0.410	0.301	*
					B	0.349	-0.149	-0.116	
					C	0.175	-0.172	-0.116	
					D	0.205	-0.000	-0.000	
					Other	0.018	-0.709	-0.237	
38	0-24	0.283	0.384	0.288	A	0.205	-0.263	-0.185	
					B	0.241	-0.089	-0.065	
					C	0.175	-0.134	-0.091	
					D	0.283	0.384	0.288	*
					Other	0.096	0.040	0.023	



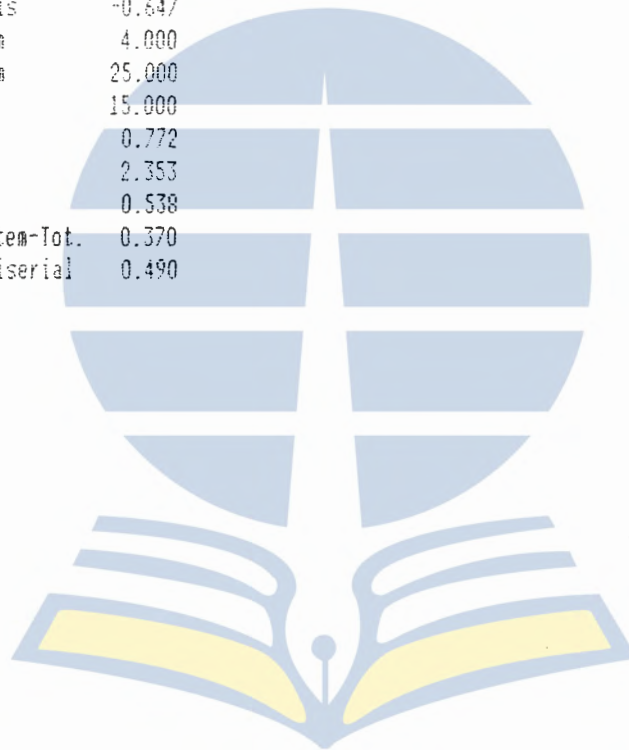
(continued)

Seq. No.	Scale -Item	Item Statistics			Alternative Statistics			
		Prop. Correct	Biser.	Point Biser.	Alt.	Prop. Endorsing	Biser.	Point Biser. Key
39	0-25	0.355	0.454	0.353	A	0.187	-0.088	-0.061
					B	0.355	0.454	0.353 *
					C	0.301	-0.209	-0.159
					D	0.114	-0.273	-0.166
					Other	0.042	-0.219	-0.098
48	0-26	0.651	0.229	0.178	A	0.096	-0.325	-0.188
					B	0.084	-0.010	-0.006
					C	0.163	-0.149	-0.099
					D	0.651	0.229	0.178 *
					Other	0.006	0.531	0.117
59	0-27	0.596	0.207	0.163	A	0.596	0.207	0.163 *
					B	0.139	-0.136	-0.087
					C	0.163	-0.069	-0.046
					D	0.048	-0.119	-0.056
					Other	0.054	-0.192	-0.093
71	0-28	0.289	0.392	0.296	A	0.289	0.392	0.296 *
					B	0.163	-0.228	-0.152
					C	0.307	-0.129	-0.098
					D	0.145	-0.125	-0.081
					Other	0.096	-0.025	-0.014
76	0-29	0.343	0.330	0.256	A	0.223	-0.122	-0.088
					B	0.120	-0.011	-0.007
					C	0.343	0.330	0.256 *
					D	0.187	-0.243	-0.167
					Other	0.127	-0.085	-0.053

(continued)

## Scale Statistics

Scale:	0
-----	
N of Items	29
N of Examinees	166
Mean	15.590
Variance	24.278
Std. Dev.	4.927
Skew	-0.038
Kurtosis	-0.647
Minimum	4.000
Maximum	25.000
Median	15.000
Alpha	0.772
SEM	2.353
Mean P	0.538
Mean Item-Tot.	0.370
Mean Biserial	0.490



## 4. MATHEMATICS IN MAY 1987 EXAMINATION

Seq. No.	Scale -Item	Item Statistics			Alternative Statistics			
		Prop. Correct	Biser.	Point Biser.	Alt.	Prop. Endorsing	Biser.	Point Biser. Key
1	0-1	0.557	0.754	0.599	A	0.158	-0.356	-0.236
					B	0.074	-0.195	-0.105
					C	0.198	-0.686	-0.479
					D	0.557	0.754	0.599 *
					Other	0.012	0.212	0.061
2	0-2	0.669	0.694	0.535	A	0.669	0.694	0.535 *
					B	0.108	-0.454	-0.272
					C	0.167	-0.389	-0.261
					D	0.053	-0.647	-0.311
					Other	0.003	-0.040	-0.007
3	0-3	0.824	0.614	0.417	A	0.087	-0.583	-0.327
					B	0.040	-0.306	-0.135
					C	0.050	-0.395	-0.186
					D	0.824	0.614	0.417 *
					Other	0.000	-9.000	-9.000
4	0-4	0.142	0.371	0.239	A	0.142	0.371	0.239 *
					B	0.350	-0.050	-0.039
					C	0.149	0.058	0.037
					D	0.353	-0.200	-0.156
					Other	0.006	-0.225	-0.050
5	0-5	0.811	0.443	0.306	A	0.115	-0.388	-0.236
					B	0.811	0.443	0.306 *
					C	0.050	-0.183	-0.086
					D	0.022	-0.518	-0.184
					Other	0.003	0.096	0.016
7	0-6	0.663	0.356	0.275	A	0.663	0.356	0.275 *
					B	0.062	-0.239	-0.121
					C	0.155	-0.277	-0.182
					D	0.121	-0.174	-0.107
					Other	0.000	-9.000	-9.000

(continued)

Seq. No.	Scale -Item	Item Statistics			Alternative Statistics			
		Prop. Correct	Biser.	Point Biser.	Alt.	Prop. Endorsing	Biser.	Point Biser. Key
8	0-7	0.864	0.685	0.436	A	0.031	-0.291	-0.117
					B	0.028	-0.561	-0.218
					C	0.864	0.685	0.436 *
					D	0.077	-0.645	-0.350
					Other	0.000	-9.000	-9.000
9	0-8	0.542	0.615	0.490	A	0.124	-0.347	-0.216
					B	0.149	-0.212	-0.138
					C	0.542	0.615	0.490 *
					D	0.180	-0.440	-0.301
					Other	0.006	-0.480	-0.107
10	0-9	0.591	0.472	0.373	A	0.170	-0.252	-0.170
					B	0.591	0.472	0.373 *
					C	0.139	-0.337	-0.216
					D	0.099	-0.255	-0.149
					Other	0.000	-9.000	-9.000
11	0-10	0.659	0.590	0.457	A	0.161	-0.171	-0.114
					B	0.659	0.590	0.457 *
					C	0.087	-0.526	-0.295
					D	0.093	-0.551	-0.315
					Other	0.000	-9.000	-9.000
12	0-11	0.793	0.734	0.518	A	0.118	-0.764	-0.468
					B	0.793	0.734	0.518 *
					C	0.034	-0.333	-0.139
					D	0.056	-0.304	-0.149
					Other	0.000	-9.000	-9.000
14	0-12	0.492	0.430	0.343	A	0.492	0.430	0.343 *
					B	0.139	-0.053	-0.034
					C	0.356	-0.406	-0.316
					D	0.012	-0.265	-0.077
					Other	0.000	-9.000	-9.000

(continued)

Seq. No.	Scale -Item	Item Statistics			Alternative Statistics				
		Prop. Correct	Biser.	Point Biser.	Alt.	Prop. Endorsing	Biser.	Point Biser.	Key
15	0-13	0.412	0.575	0.455	A	0.297	-0.107	-0.081	
					B	0.167	-0.473	-0.317	
					C	0.412	0.575	0.455	*
					D	0.121	-0.326	-0.201	
					Other	0.003	-0.311	-0.053	
16	0-14	0.310	0.574	0.438	A	0.211	-0.039	-0.028	
					B	0.291	-0.260	-0.197	
					C	0.183	-0.353	-0.242	
					D	0.310	0.574	0.438	*
					Other	0.006	-0.480	-0.107	
17	0-15	0.455	0.365	0.291	A	0.282	-0.238	-0.179	
					B	0.455	0.365	0.291	*
					C	0.214	-0.134	-0.095	
					D	0.031	-0.209	-0.084	
					Other	0.019	-0.232	-0.078	
18	0-16	0.656	0.586	0.454	A	0.136	-0.227	-0.145	
					B	0.124	-0.478	-0.297	
					C	0.080	-0.415	-0.228	
					D	0.656	0.586	0.454	*
					Other	0.003	-0.649	-0.110	
19	0-17	0.322	0.317	0.243	A	0.142	-0.255	-0.164	
					B	0.322	0.317	0.243	*
					C	0.201	-0.145	-0.102	
					D	0.300	-0.066	-0.050	
					Other	0.034	0.098	0.041	
20	0-18	0.440	0.540	0.429	A	0.254	-0.193	-0.142	
					B	0.195	-0.321	-0.223	
					C	0.440	0.540	0.429	*
					D	0.096	-0.296	-0.171	
					Other	0.015	-0.310	-0.098	

(continued)

Seq. No.	Scale -Item	Item Statistics			Alternative Statistics			
		Prop. Correct	Biser.	Point Biser.	Alt.	Prop. Endorsing	Biser.	Point Biser. Key
25	0-19	0.848	0.417	0.273	A	0.059	-0.224	-0.111
					B	0.059	-0.425	-0.212
					C	0.031	-0.227	-0.091
					D	0.848	0.417	0.273 *
					Other	0.003	-0.649	-0.110
27	0-20	0.341	0.304	0.235	A	0.341	0.304	0.235 *
					B	0.108	-0.152	-0.091
					C	0.415	-0.030	-0.024
					D	0.130	-0.297	-0.187
					Other	0.006	-0.480	-0.107
28	0-21	0.316	0.483	0.369	A	0.133	-0.253	-0.160
					B	0.303	-0.271	-0.206
					C	0.241	-0.077	-0.056
					D	0.316	0.483	0.369 *
					Other	0.006	0.067	0.015
31	0-22	0.279	0.329	0.247	A	0.276	0.053	0.040
					B	0.276	-0.229	-0.171
					C	0.279	0.329	0.247 *
					D	0.139	-0.225	-0.144
					Other	0.031	-0.026	-0.010
35	0-23	0.381	0.393	0.309	A	0.381	0.393	0.309 *
					B	0.368	-0.097	-0.076
					C	0.146	-0.336	-0.218
					D	0.096	-0.158	-0.091
					Other	0.009	-0.377	-0.098
38	0-24	0.526	0.653	0.520	A	0.155	-0.003	-0.002
					B	0.214	-0.562	-0.399
					C	0.105	-0.525	-0.311
					D	0.526	0.653	0.520 *
					Other	0.000	-9.000	-9.000

(continued)

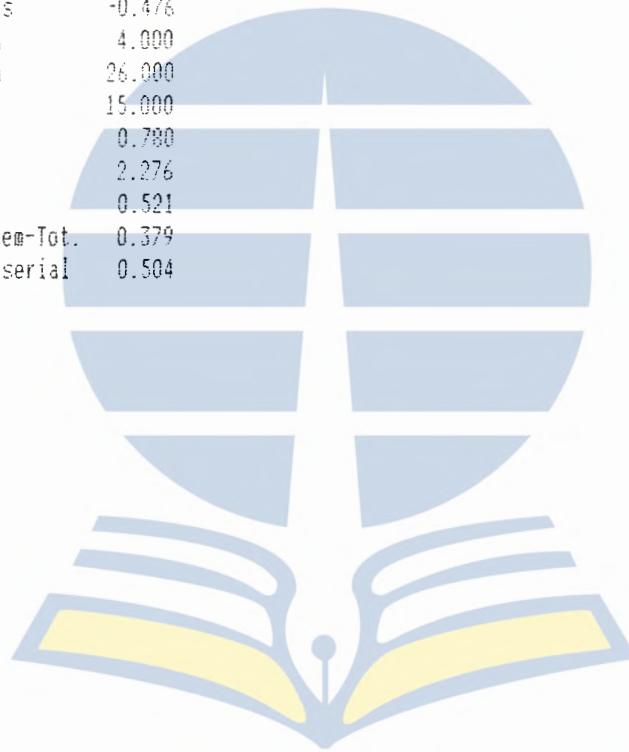
Seq. No.	Scale	Correct	Prop.	Item Statistics					Alternative Statistics					
				Point	Biserr.	Biserr.	Alt.	Endorsing	Prop.	Point	Biserr.	Biserr.	Key	
39	0-25	0.743	0.517	0.382	A	0.056	-0.446	-0.218	-0.443	-0.246	B	0.084	-0.443	-0.246
					C	0.743	0.517	0.382	*	0.382				
					D	0.108	-0.242	-0.145						
					Other	0.009	-0.147	-0.038						
41	0-26	0.396	0.412	0.325	A	0.406	-0.228	-0.180						
					B	0.127	-0.015	-0.010						
					C	0.065	-0.504	-0.259						
					D	0.396	0.412	0.325	*	0.325				
					Other	0.006	-0.189	-0.042						
42	0-27	0.263	0.401	0.297	A	0.303	-0.079	-0.060						
					B	0.173	-0.087	-0.059						
					C	0.257	-0.244	-0.180						
					D	0.263	0.401	0.297	*	0.297				
					Other	0.003	-0.243	-0.041						
44	0-28	0.291	0.491	0.371	A	0.158	-0.161	-0.106						
					B	0.204	-0.325	-0.229						
					C	0.341	-0.093	-0.072						
					D	0.291	0.491	0.371	*	0.371				
					Other	0.006	-0.189	-0.042						

(continued)

## Scale Statistics

Scale: 0  
-----

N of Items	28
N of Examinees	323
Mean	14.585
Variance	23.574
Std. Dev.	4.855
Skew	-0.095
Kurtosis	-0.476
Minimum	4.000
Maximum	26.000
Median	15.000
Alpha	0.780
SEM	2.276
Mean P	0.521
Mean Item-Tot.	0.379
Mean Biserial	0.504





Appendix E

Results of Item Analyses using the Rasch Model



## 1. ENGLISH IN DECEMBER 1986 EXAMINATION

Item	Difficulty	Chi Sq.	df
1	-0.637	8.592	8
6	1.138	9.501	8
11	1.138	14.685	8
13	-0.084	8.110	8
16	-0.637	13.697	8
17	0.027	15.109	8
18	-0.637	6.293	8
19	0.370	10.055	8
26	0.998	9.979	8
27	-0.525	10.050	8
32	-0.414	5.115	8
33	-1.103	6.493	8
35	0.140	8.258	8
37	0.864	7.556	8
38	-2.375	14.193	8
42	0.735	7.592	8
44	-0.864	7.909	8
49	-0.982	6.699	8
50	-0.084	3.113	8
52	0.140	5.067	8
58	0.027	6.937	8
59	-0.414	7.619	8
60	0.370	15.883	8
61	1.965	4.373	8
62	-0.084	8.180	8
63	0.254	10.413	8
65	0.027	20.055	8
69	0.140	11.597	8
70	0.140	5.615	8
78	0.370	5.869	8

## 2. ENGLISH IN MAY 1987 EXAMINATION

Item	Difficulty	Chi Sq.	df
2	-1.286	21.880	14
3	-0.290	8.792	14
6	-1.650	8.087	14
7	0.368	15.163	14
8	0.156	19.174	14
9	-0.614	18.632	14
10	-0.213	64.230	14
11	1.594	12.498	14
13	-0.137	12.896	14
15	-1.940	15.315	14
16	0.227	13.230	14
17	-0.530	11.212	14
19	-0.213	20.442	14
24	-0.137	8.584	14
30	0.717	22.887	14
31	0.787	15.426	14
32	0.156	14.209	14
33	0.857	7.597	14
35	-0.614	9.468	14
36	-0.063	17.724	14
38	-0.137	12.627	14
39	-0.290	25.121	14
40	-0.213	15.602	14
42	0.227	16.758	14
43	-1.400	5.425	14
46	0.787	14.264	14
48	-0.530	10.217	14
49	-0.448	21.045	14
50	-0.976	9.922	14
51	-0.448	13.499	14
53	0.368	9.338	14
54	0.787	16.780	14
55	-0.137	11.568	14
57	-0.290	7.511	14
58	-0.063	12.562	14
61	0.227	9.987	14

(continued)

<u>Item</u>	<u>Difficulty</u>	<u>Chi Sq.</u>	<u>df</u>
62	0.084	7.156	14
67	1.756	10.155	14
69	0.368	11.702	14
73	0.368	10.787	14
76	0.857	35.700	14
77	0.998	19.848	14
78	0.927	11.934	14



## 3. MATHEMATICS IN DECEMBER 1986 EXAMINATION

Item	Difficulty	Chi Sq.	df
1	0.490	13.839	16
2	0.020	7.266	16
3	0.575	11.286	16
5	-0.678	25.976	16
6	-2.407	8.357	16
7	-1.082	14.197	16
8	-0.616	28.223	16
9	-0.232	22.862	16
11	1.045	7.624	16
12	-1.394	19.590	16
14	0.351	11.577	16
15	-0.710	17.775	16
17	1.171	33.891	16
19	0.324	18.686	16
20	0.804	24.444	16
21	-1.272	12.375	16
23	-1.394	6.647	16
25	-0.941	15.359	16
26	-0.176	19.247	16
27	-0.063	14.217	16
28	-0.119	21.735	16
30	1.237	23.766	16
36	1.441	31.339	16
38	1.270	16.995	16
39	0.893	16.596	16
48	-0.494	37.057	16
59	-0.232	34.954	16
71	1.237	25.296	16
76	0.953	34.140	16

## 4. MATHEMATICS IN MAY 1987 EXAMINATION

Item	Difficulty	Chi Sq.	df
1	-0.137	27.151	18
2	-0.683	16.762	18
3	-1.629	23.533	18
4	2.156	15.829	18
5	-1.537	24.836	18
7	-0.651	28.263	18
8	-1.965	11.082	18
9	-0.065	15.292	18
10	-0.299	22.805	18
11	-0.635	14.476	18
12	-1.407	21.416	18
14	0.165	20.515	18
15	0.541	7.918	18
16	1.051	17.139	18
17	0.337	25.712	18
18	-0.619	32.045	18
19	0.986	43.301	18
20	0.410	9.540	18
25	-1.828	12.925	18
27	0.890	51.017	18
28	1.018	23.577	18
31	1.220	24.476	18
35	0.690	26.348	18
38	0.007	12.785	18
39	-1.093	5.567	18
41	0.615	35.564	18
42	1.308	23.433	18
44	1.151	34.357	18

## Appendix F

### Questions to be Answered in Designing Item Banking Systems



## I. ITEMS:

### A. Acquisition & Development.

1. Develop/use your own item collection or use collections of others?
  - a. if develop your own item collection what development procedures will be followed?
  - b. If use collections of others, will the items be leased or purchased, and is the classification scheme sufficiently documented and the item format specifications sufficiently compatible for easy transfer and use?
2. What types of "item" will be permitted?
  - a. Will open-ended (constructed response) items, opinion questions, instructional objectives, or descriptions of performance tasks be included in the bank?
  - b. Will all the items be made to fit a common format (e.g. all multiple choice with options a,b,c,and d)?
  - c. Must the items be calibrated, validated, or otherwise carry additional information?
3. What will be the size of the item collection?
  - a. How many items per objective/subtopic (collection depth)?
  - b. How many different topics (collection breadth)?
4. What review, tryout and editing procedures will be used?
  - a. Who will perform the review/editing?
  - b. Will there be a field tryout, and if so, what statistics will be gathered, and what criteria will be used for inclusion into the bank?

### B. Classification

1. How will the subject matter classifications be conducted?
  - a. Will the classification by subject matter use fixed categories, keywords, or some combination of the two?
  - b. Who will be responsible for preparing the taxonomy?
  - c. How detailed will the taxonomy be? Will it be hierarchically or non-hierarchically arranged?



- d. Who will assign classification indices to each item, and how will this assignment be verified?
2. What other assigned information about the items will be stored in the item bank? (See attached list for potential attributes).
3. What measured information about the items will be stored in the bank? (See Appendix B list for potential measures). How will the item measures be calculated?

### C. Management

1. Will provision be made for updating the classification scheme and items? If so:
  - a. Who will be permitted to make additions, deletions, and revisions?
  - b. What review procedures will be followed?
  - c. How will the changes be disseminated?
  - d. How will duplicate (or near duplicate) items be detected and eliminated?
  - e. When will a revision of an item be trivial enough that items statistics from the current version?
  - f. Will item statistics be stored from each use, last use, or aggregated across uses?
2. How will items that require pictures, graphs, special characters, or other types of enhanced printing be handled?
3. How will items that must accompany other items, such as a series of questions about the same reading passage, be handled?

## II. TEST

### A. Assembly

1. Must the test constructor specify the specific items to appear on the test or will the items be selected by the computer?

2. If the items are selected by the computer, :
    - a. How will one item out of several that matches the search specification be selected (randomly, time since, last usage, frequency or previous use)?
    - b. What happens if no item meets the search specifications?
    - c. Will a test constructor have the option to reject a selected item, and if so, what will be the mechanism for doing so?
    - d. What precautions will be taken to insure that examiners who are tested more than once do not receive the same items?
  3. What item or test parameters can be specified for test assemble (item format restrictions, limits on difficulty levels, expected score distribution, expected test reliability, etc.)?
  4. What assembly procedures will be available (options to multiple choice items placed in random order, the test items placed in random order, different on each test)?
  5. Will the system print test or just specify which items to use? If the former, how will the tests be printed or duplicated and where will the answers be displayed?
- B. Administration, Scoring and Reporting
1. Will the system be capable of on-line test administration? If so:
    - a. How will access be managed?
    - b. Will test administration be adaptive, and if so, using what procedures?
  2. Will the system provide for test scoring? If so:
    - a. What scoring formula will be used (right only, correction for guessing, partial credit for some answers, weighting by discrimination values)?

- b. How will constructed response be evaluated (off-line by the instructor, on-line/off-line by examiners composing their answers to a key, on-line by computer with/without employing a spelling algorithm)?
3. Will the system provide for test reporting? If so:
  - a. What records will be kept (the tests themselves, individual student item responses, individual student test scores, school or other group scores) and for how long? Will new scores for individuals, and groups supplement or replace old scores?
  - b. What reporting options (contents/format) will be available?
  - c. To whom will the reports be sent?

#### C. Evaluation

1. Will reliability and validity data be collected? If so, what data will be collected by whom, and how will they be used?
2. Will norms be made available and, if so, based on what norm-referenced measures?

### III. SYSTEM:

#### A. Acquisition and Development

1. Who will be responsible for acquisition/development, given what resources, and operating under what constraints?
2. Will the system be made transportable to others? What levels and what degree of documentation will be available?

#### B. Software/Hardware Features

1. What aspects of the system will be computer assisted?
  - a. Where will the items be stored (computer, paper, card file)?
  - b. Will request be filled using a batch, on-line, or manual mode?

2. Will items be stored as on large collection or will separate files be maintained for each user?
3. How will the item banking system be constructed (from scratch; by piecing together word processing, database management, and other general purpose programs; by adopting existing item banking systems)?
4. What specific equipment will be needed (for storage, retrieval, interactions with the system, etc.)?
5. How user and maintenance friendly will the equipment and support programs be?
6. Who will be responsible for equipment maintenance?

#### C. Monitoring and Training

1. What system features will be monitored (number of items per classification category, usage by user group, number of revisions until a user is satisfied, distribution of test lengths or other test characteristics, etc.)
2. Who will monitor the system, train users, and give support (initially, ongoing)?
3. How will information about changes in system procedures be disseminated?

#### D. Access and Security

1. Who will have access to the items and other information in the bank (authors/owners, teachers, students)? Who can request tests?
2. Will users have direct access to the system or must they go through an intermediary?
3. What procedures will be followed to secure the contents of the item bank (if they are to be secure)?
4. Where will the contents of the item bank be housed (centrally or will each user also have a copy)?

5. Who will have access to score reports?

#### IV. USE AND ACCEPTANCE

##### A. General

1. Who decides to what uses the item bank will be put? And will these uses be the ones that the test users need and want?
2. Who will develop the tests and who will be allowed to use the system? Will these people be acceptable to the examines and recipient of the test information?
3. Will the system be able to handle the expected demand for use?
4. Will the output of the system likely to be used and used as intended?
5. How will user acceptable and item bank credibility be enhanced?

##### B. Instructional Improvement. If This is an intended use:

1. Will the item bank be parts of larger instructional/decision making system?
2. Which textbooks, curriculum guidelines, and other materials, if any, will be keyed to the bank's items? Who will make that decision and how will the assignments be validated?
3. Will items be available for drill and practice as well as for testing?
4. Will information be available to users that will assist in the diagnosis of educational needs?

##### C. Adaptive Testing. If this is an option:

1. How will the scheduling of the test administrations take place?

2. How will the items be selected to measure testing efficiency yet maintain content representation and avoid duplication between successive test administrations?
3. What criteria will be used to terminate testing?
4. What scoring procedures will be followed?

D. Certification of Competence. If this is an intended use:

1. Will the item bank contain measures that cover all the important component skills of the competence being assessed?
2. How many attempts at passing the test will be allowed: When? How will these attempts be monitored?

E. Program/Curriculum Evaluation. If these is an intended use:

1. Will it be possible to implement the system so as to provide reliable measures of student achievement in a large number of specific performance areas?
2. Will the item bank contain measures that covers all the important stated objectives of the curriculum? That go beyond the stated objectives of the curriculum?
3. Will the item bank yield commensurable data that permit valid comparisons over time?

F. Testing and Reporting Requirements Imposed by External Agencies. If meeting these requirements is an intended use:

1. Will the system be able to handle requirements for program evaluation (e.g. Ch. I), student selection into specially funded programs, assessing educational needs, and reporting?
2. Will the system be able to accommodate minor modifications in the testing and reporting requirements?

## V. COSTS

### A. Cost Feasibility

1. What are the (fixed, variable) costs (financial, time, space, equipment and supplies) to create and support the system?
2. Are these costs affordable?

### B. Cost Comparisons

1. How do the item banking system costs compare to the present or other testing system that achieve the same pools?
2. Do any expanded capabilities justify the extra cost? Are any restricted capabilities balanced by cost savings?



### References

- Anderson, J. (1985). Item Banking at UT: A Discussion Paper and Preliminary Proposal. Paper, Unpublished proposal for Universitas Terbuka.
- Arter, J.A., & Estes, G.D. (1985). Item Banking for Local Test Development: Practitioner's handbook. Portland, OR: Northwest Regional Educational Lab. (ERIC Document Reproduction Service No. ED 266 166).
- Assessment System Corporation. (1986). User's Manual for Iteman, Rascal, and Ascal. St. Paul, MA: Assessment System Corporation.
- Bejar, I.I., Weis, D.J., & Kingsbury, G.G. (1977). Calibration of an Item Pool for the Adaptive Measurement of Achievement. Minneapolis, MN: Psychometric Methods Program, Department of Psychology, University of Minnesota. (ERIC Document Reproduction Service No. ED 146 231).
- Burke, N.W., Kaufman, B.D., & Webb, N. (1985). The Wisconsin Item Bank. In G.D. Estes (Ed.). Examples of Item Bank to Support Local Test Development: Two case studies with reactions. Portland, OR: Evaluation and Assessment North West Regional Educational Library. (ERIC Document Reproduction Service No. ED 266 168).
- Choppin, B. (1978). Item Banking and the Monitoring of Achievement: An introductory paper. National Foundation for Educational Research in England and Wales. England.
- Dennis, D., Nickel, P., & Estes, G. (1985). Review of Microcomputer Item Banking Software. Portland, OR: Evaluation and Assessment Northwest Regional Educational Library. (ERIC Document Reproduction Service No. ED 266 167)
- Departemen Pendidikan dan Kebudayaan. (1984). Initial Planning Consideration. (Unpublished).



Gronlund, N.E. (1985). Measurement and Evaluation in Teaching (5th Ed.). New York: MacMillan.

✓ Hambleton, R.K. (1979). Latent Trait Models and their Applications. New Directions for Testing and Measurement 4, 13-32.

Hambleton, R.K. (1983). Application of Item Response Theory. Vancouver, BC: Educational Research Institute of British Columbia.

Hambleton, R.K. & Swaminathan, H. (1985). Item Response Theory: Principles and Applications. Hingham, MA: Kluwer Boston.

Hambleton, R.K. et. Al. (1978). Developments in Latent Trait Theory: Models, Technical Issues, and Application. Review of Educational Research, 48.(4), 467-510.

Holmberg, B. (1982). Distance Education: A Short Handbook. Stockholm: Liber Hermonds.

Hulin, C.L., Drasgow, F., & Parsons, C.H. (1983). Item Response Theory: Application to psychological measurement. Homewood, IL: Dow Jones-Irwin.

Kaplan, R.M., & Saccuzzo, D.P. (1982). Psychological Testing: Principle, applications, and issues. Monterey, CA: Brooks/Cole.

Keegan, D.J. (1983). On defining distance education. In Seward, D., Keegan, D.J. & Holmberg, B. (ed.) Distance Education: International perspectives. NY: St. Martin's.

Mead, R.J. (1981). Basic Ideas in Item Banking. Los Angeles, CA: MCME. (ERIC Document Reproduction Service No. ED 208029).

Millman, J., & Arter, J.A. (1984). Issues in Item Banking. Journal of Educational Measurement, 21(4), 315 - 330

- Ministry of Education and Culture. (1984). Information Booklet on Universitas Terbuka. Jakarta: Universitas Terbuka.
- O'Brien, ME & Tohn, D. (1984). Applying and Evaluating Rasch Vertical Equating Procedures for out-of-level Testing. West Palm Beach, FL: Paper presented at the annual meeting of Eastern Educational Research
- Perraton, H. (1983). A Theory for Distance Education. In Sewart, D., Keegan, D., & Holmberg, B. (Eds.). Distance Education: International Perspectives. NY: St. Martin's.
- Popham, W.J. (1981). Modern Educational Measurement. Englewood Cliffs, NJ: Prentice-Hall.
- Rentz, R.R., & Rentz, C.C. (1978). Does the Rasch Model Really Work? A discussion for practitioners. Princeton, NJ: Educational Testing Service.
- Robitaille, D.F., & O'Shea, T. (1983). The Development of an Item Bank in Mathematics Using the Rasch Model. Canadian Journal of Education 8(1), 57-70.
- Sangat Jelek Ujian Sekolah dengan "Multiple Choice". (1986, September). Kedaulatan Rakyat. p.1
- Sewart D., Keegan, D., & Holmberg, B. (Eds.). (1983). Distance Education: International Perspectives. NY: St. Martin's.
- Willmott, A.S., & Fowles, D.E. (1974). The Objective Interpretation of Test Performance: The Rasch model applied. Berk, England: NFER.
- Wisniewski, D.R. (1986). An Application of the Rasch Model to Computerized Adaptive Testing. San Francisco, CA: Paper presented at the 70th Annual Meeting of the American Educational Research Association. (ERIC Document Reproduction Service No. ED 268 183).

Woodley, A. (1979). Institutional Performance How Open is open. In D. Billing, Indicators of Performance. (ERIC Document Reproduction Service No. ED 223 137).

Wright, B.D. (1977) Solving Measurement Problems with the Rasch Model. Journal of Education Measurement, 4, (2), 97-116.

Wright, B.D. & Stone, M.H. (1979). Best Test Design. Chicago, IL: MESA

