ACCEPTED [redacted] DEAN 28 Apr 93

Comparison of Test Items Within and Between Faculties at The Indonesian Open University

by

Sondang Purnamasari Pakpahan
B.Sc., University of Indonesia, 1987

A Thesis Submitted in Partial Fulfillment of
the Requirements for the Degree of

MASTER OF ARTS

in the Department of
Psychological Foundations in Education

We accept this thesis as conforming
to the required standard

[redacted]

Dr. Daniel G. Bachor, Supervisor
(Department of Psychological Foundations)

[redacted]

Dr. John O. Anderson, Departmental Member
(Department of Psychological Foundations)

[redacted]

Dr. Laurence E. Devlin, Outside Member
(Department of Communication and Social Foundations)

[redacted]

Dr. Leslee G. Francis, External Examiner
(Department of Social and Natural Sciences)

Supervisor: Dr. Daniel G. Bachor

ABSTRACT

The purpose of this study was to investigate the differences in the quality of UT's (Universitas Terbuka's) test items in terms of item difficulty, discrimination, and test reliability, using the criteria proposed by Ebel and UT. A comparison was made between test items used previously and a sample of revised items drawn from the same sources. This investigation was conducted in two faculties, the Faculty of Education (FKIP) and the Faculty of Mathematics and Natural Science (FMIPA).

This study involved test results of three Physics and three English courses from FKIP, and two Mathematics and three Statistics courses from FMIPA drawn from three test administrations. The new tests consisted of revised and non-revised items drawn from the two old test administrations.

The major finding was that most of the test items in FKIP and FMIPA met UT's standards. FKIP and FMIPA had the same approximate qualities in item difficulty. However, the qualities of item discrimination and test reliability, were slightly better in FKIP. Most of the items in both faculties are still classified as

'hard' with marginal discrimination. Most of the tests in both faculties were also more difficult, less discriminating, and less reliable than testing experts posit as ideal, especially for science courses such as Physics, Mathematics, and Statistics.

There were three other major findings. First, item revisions in three of the six FKIP courses (Introduction to Quantum Mechanics, English for Arts and Science, and English for Education) resulted in desired changes in both the item difficulty and discrimination. Item revisions in the other courses, Alternating Current, only met the desired criterion for item discrimination. Second, item revisions in two of the five FMIPA courses (Calculus I and Survey Sampling Method) resulted in desired changes in both the item difficulty and item discrimination. Item revisions in the other course, Applied Experimental Design, only resulted in desired changes in item discrimination. Third, overall, a large percentages of revised and non-revised items in FKIP and FMIPA still had low discriminations (less than 0.30), implying that further item revision should be conducted in both FKIP and FMIPA.

Examiners:

[redacted]

Dr. Daniel G. Bachor, Supervisor
(Department of Psychological Foundations)

[redacted]

Dr. John O. Anderson, Departmental Member
(Department of Psychological Foundations)

[redacted]

Dr. Laurence E. Devlin, Outside Member
(Department of Communication and Social Foundations)

[redacted]

Dr. Leslee G. Francis, External Examiner
(Department of Social and Natural Sciences)

TABLE OF CONTENTS

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Dr. Daniel G. Bachor, and the other members of my committee, Dr. John O. Anderson and Dr. Laurence E. Devlin, for their support, direction, and encouragement.

I would also like acknowledge, Carol M. Harvey, who has assisted me in improving the grammar.

A special thank you to my friends in the Faculty of Mathematics and Natural Science and the Faculty of Education at the Indonesian Open University who have assisted me in collecting the data.

A very special thank you to my friends who have inspired, encouraged, and supported me in my work.

CHAPTER I

Introduction

The Importance of Test Evaluation

Tests serve a variety of functions. Tests can be used for selection, as when new employees are selected from a group of job applicants; for classification, as when a person is classified as weak in verbal skills and strong in mechanical skills; and for evaluation, as when students are assigned grades in a class. Tests can also be useful in counseling and research (Allen & Yen, 1979; Ebel & Frisbie, 1986).

In this thesis, the role testing plays in distance education was considered. The need to evaluate the quality of tests provided in one distance education setting was considered.

To date, the use of achievement tests in learning is widespread throughout the world (Ebel & Frisbie, 1986). Many investigators have examined the appropriateness of achievement tests and discussed such topics as the following four: (1) the problems in determining the optimal values of item difficulty, discrimination, and test reliability (Ebel, 1972; Ebel & Frisbie, 1986; Englehart, 1965; Henryssen, 1971; Kelley, 1939; Lien, 1976; Lord & Novick 1968; Pyrczak, 1973); (2) the problems in identification of item analysis

techniques to evaluate the quality and usefulness of test items (Davis, 1952; Sato, 1980; Turnbull, 1956); (3) the effects of item formats on item difficulty, discrimination, and test reliability (Dudycha & Carpenter, 1973; Lancaster, 1987; Kolstad, 1991; Maihoff & Mehrens, 1985; Tollefson, 1987); and (4) the effects of item order on item difficulty, discrimination, and test reliability (Allison, 1984; Balch, 1989; Bresnock, 1989; Chissom & Chukabarah, 1985; Hambleton & Traub, 1974; Huck & Bowers, 1972; Plake, 1980; Plake et al., 1982; Sander, 1988). However, in examining student achievement, the issue of evaluating test quality using item analysis in a distance education setting has received minimal attention. McMillan, Mundrake, and McGuire (1989) did evaluate the quality of tests at a business school based on item difficulty and item discrimination. In discussing such studies, Ebel and Frisbie (1986, 1991) provided two reasons for analyzing and evaluating a test. First, teachers should assess the quality of the tools of assessment. If students perform below expectation, perhaps the explanation lies in the methods of evaluation rather than in inadequate learning on the students' part. Tests should be evaluated, then, to determine if the scores they yield have value for the purpose for which they were

originally intended. Second, analyzing test and item data can reveal technical flaws and errors of judgment made by item writers. By making use of these data, the writers can improve their item-writing skills, and at the same time, revise their test items for future use. Loftus (1988) underlined the importance of studying the quality of tests in a distance education setting. He stated that recent studies on home study course completion rates have pointed out the critical importance of well-constructed examinations. At the Indonesian Open University, Universitas Terbuka (UT), an examination of the quality of recent revisions in test items for instructional modules has not yet been conducted.

Achievement Tests in Distance Education

Achievement tests play important roles in distance education. First, the tests serve functions in reinforcing students' learning. The students in distance education setting, who rely heavily on written course materials, frequently tend to forget much of what they have read over the duration of the course. As students face the challenge of tests, they review and perhaps do extra work to prepare themselves to write the tests. Second, the tests serve two other functions. Test results indicate to the institution which students

passed the tests and thus qualify to take further courses. At the same time, these results provide some indication of the effectiveness of the instructional program (MacKenzie, Christensen, & Rigby, 1968).

Most correspondence study educators consider the multiple choice test to be the best measurement tool (Foltz, 1990; Loftus, 1988; MacKenzie, Christensen, & Rigby, 1968). MacKenzie, Christensen, and Rigby (1968) argued that, as subjective responses are difficult and costly to correct; therefore, correspondence programs frequently resort to objective testing. Loftus (1988) also indicated that most home study courses make heavy use of the multiple choice examination. Foltz (1990) stated that the multiple choice format serves home study education needs best. According to Foltz, the multiple choice test can be graded quickly and economically.

Achievement Tests at Universitas Terbuka

At the Indonesian Open University, Universitas Terbuka (UT), which offers a distance learning programs, tests are conducted twice a year at the end of each semester. Tests administered in the first semester of 1992, for example, are called 92.1 tests, while those administered in the second semester of 1992 are called 92.2 tests. The tests are conducted in 90-minute periods and most of these tests are written in four-

option multiple choice format. The tests consist of 20 to 40 items for Mathematics and Science courses, and 60 to 90 items for Social Science courses. Due to the large number of UT students spread throughout all provinces in Indonesia, multiple choice tests are used for most of UT's courses.

UT's testing methods have tended toward norm-referenced measurement. This system is believed to be more appropriate to UT since the instructional objectives stated in UT's instructional materials (modules) and Tables of Specifications, in most cases are general objectives, not specific enough to be used for criterion-referenced test development. Some of the Tables of Specifications, the Table of Specifications for Calculus I (see Appendix F); for example; stated only the topics and sub-topics of the instructional materials. This problem obviously has led to difficulties in test item writing where subjective interpretations have occurred. Therefore, it is very doubtful that good criterion-referenced test items can be produced in this situation.

In detail, the test development and test construction procedures at UT can be divided into 6 steps. The procedure begins with the writing of a Table of Specifications. The Table of Specifications at UT

consists of topics or objectives and the cognitive domain of the items to be written. The topics and objectives in the Table of Specifications are similar to what is stated in the instructional materials (modules). However, many of the modules state only general objectives. Therefore, many developers of Table of Specifications prefer to write specific objectives based on the general objectives in the modules.

Classification of items in the cognitive domain is based on Bloom's taxonomy (Bloom, 1956). At UT, not all faculties use a Table of Specifications in developing their tests. Only the Faculty of Mathematics and Natural Science (FMIPA) and the Faculty of Education (FKIP) use this device. The other faculties, the Faculty of Economics (FEKON) and the Faculty of Social and Political Science (FISIP), do not use Tables of Specifications in developing their tests. In FEKON, although Tables of Specifications are available, the test writers do not use these devices in developing their tests. The test writers did not use the Tables of Specifications because they did not develop them and/or are not in agreement with the content reflected in the tables. Since most of items do not follow the Table of Specifications, the staff of FEKON find it difficult when assembling items on the basis of Table of

Specifications. Therefore, the item assembly process also does not follow the Table of Specifications.

In FISIP, on the other hand, most of courses do not have Tables of Specifications for their tests. The test writers feel they do not need to develop Tables of Specifications in writing tests because they are the content experts; they are the module writers; and they are experienced in test writing. Therefore, they argue that they know what objectives they should measure. At the present time, however, FISIP is in the process of developing the Table of Specifications. FISIP encourages the test writers to develop the Tables of Specifications.

The second step is item writing. At UT, item writing is conducted every semester. Every semester, UT asks its test writers to build new test items which will be offered to the students. This policy is still in place since the item bank has not yet been fully utilized. However, in special cases where the test writer can not complete the items in the determined time, UT asks its staff to assemble items from the available item bank.

UT's test writers are content experts from traditional universities throughout Indonesia. However, some of UT's faculties, such as the Faculty of Education

(FKIP) and the Faculty of Politics and Social Science (FISIP), also ask those members of their staff who have relevant academic backgrounds and who have been involved in test construction training to write test items. In the first year of UT's establishment, 1984, the test writers were involved in test construction training. Test construction training is still conducted when UT requests new traditional university professors to be test writers.

The third step in the test construction procedure is to review the test items. Other content experts or UT's staff members who have relevant academic backgrounds and who have been trained in test construction conduct this review by examining items logically. The criteria used are the representativeness of items in relation to the given objectives, the clarity of the stem, the effectiveness of the distractors, and the correctness of the key option. The reviewers decide which items are to be accepted and which are to be revised.

The fourth step is administering the test. Sets of tests which have been prepared and printed are distributed to the regional centers. UT's regional centers organize and supervise the examinations, then collect the answer sheets and return them to UT. The

examinations are conducted in places selected by UT, such as junior and senior high schools.

The fifth step in the procedure is item analysis. Usually, pilot testing of items precedes test use and item analysis. The results are very important in helping the test constructors select the test items. Unfortunately, pilot testing is not done at UT. Item analysis is done using data of students' responses from the actual examination, after the test has been administered. Since test items are not pilot-tested, it may be that some items do not work as intended. Regarding this situation, item analysis could be used to analyze the quality of UT's test and test items and to delete low quality items from the test.

Item analysis at UT is based upon norm-referenced measurement using a combination of item difficulty, item discrimination, and test reliability. UT considers three levels of items: very good, good, and poor items. Very good items should have high discrimination (0.40 and up) and good items should have discrimination in a range of 0.20 to 0.39. The items with discrimination lower than 0.20 are considered to be poor items (Pantap coordination meeting, July 24, 1987). These criteria are slightly different from Ebel's criteria (see Table 1). Ebel (1972) proposed four levels of items based on

Table 1
UT's and Ebel's Criteria for Item Discrimination

UT		Ebel	
Index of discrimination	Item evaluation	Index of discrimination	Item evalution
0.40 and higher	Very good items	0.40 and higher	Very good items
0.20 tO 0.39	Good items	0.30 to 0.39	Reasonably good items
		0.20 to 0.29	Marginal items
Below 0.20	Poor items	Below 0.20	Poor items

the values of the discrimination as follows: 0.40 and up as very good items; 0.30 to 0.39 as reasonably good, possibly subject to improvement; 0.20 to 0.29 as marginal items with considerable need for improvement; and those below 0.20 are poor items which should be rejected or improved by revision. UT accepts discrimination of 0.20 and up in order to have wider leeway in selecting items. With these criteria, UT includes marginal items, and because of the scarcity of good quality items, this condition is to be tolerated.

In terms of item difficulty, UT uses evaluation criteria as follows: less than 0.30 indicates that an item is very difficult; 0.30 to 0.40, difficult; 0.41 to 0.84, moderate; 0.85 to 0.90, easy; and greater than 0.90, very easy. The acceptable difficulty level for UT's test items are in the range of 0.20 to 0.90 (Pantap coordination meeting, July 24, 1987). This is a wide

range of difficulty levels, and both relatively easy and very difficult items are included. The decision to accept the wider range of difficulty levels was made because the number of good items available is very limited.

UT uses criteria for the values of reliability of tests as follows: greater than 0.80 are strong reliabilities, 0.60 to 0.80 are moderate reliabilities, and those below 0.60 are weak reliabilities (Pantap coordination meeting, July 24, 1987).

The sixth and final step in the test construction procedure is item banking. The item banking system at UT is still in the development process, a process which began in February, 1985. The original purpose of item banking was to store the test items. To make UT's item bank more useful, UT began in 1989 to use item analysis as the basis of revising items for future test item use. The advantage of item analysis in revising items is supported by Lange, Lehmann, and Mehrens (1967). They argued that items can be improved, without too much effort, through use of a complete item analysis. They also suggested that revising items may be a more economical process for obtaining good items for future tests than discarding those items and replacing them with new items.

UT has revised all of the items which were offered by test administration 89.2. UT revised the test items in two ways based on the item discrimination, item difficulty, and clarity of the item. First, UT asked the reviser to analyze the item analysis, and to examine closely the extremely difficult items and items with negative discrimination indices. The extremely difficult items and the items with negative discrimination indices were examined in order to determine if they were poorly written, if they possibly represented content unfamiliar to the students, or if they represented incorrect information. If the item was too difficult (the p-value was less than 0.10), UT asked the reviser to make the stem more clear by revising the structure and adding some explanation. On the other hand, if the item was too easy (the p-value was greater than 0.90), UT asked the reviser to delete some explanation which made the item too easy and to improve the homogeneity of options. If the item discrimination was too low (the value was less than 0.30), UT asked that the options which had positive or high discriminations be revised (Zainul, 1992). Second, UT asked the reviser to review the items by paying attention to the clarity of the items, the correctness of the options, the appropriateness of the difficulty

level, the consistency of the items and Table of Specifications, and the consistency of notations used in the tests and in written course materials.

The revision activity produced three kinds of items: non-revised items, revised items, and unacceptable items. An item was not revised and called a non-revised item if the item fulfilled the following criteria: first, based on experts' judgment, the item was consistent with the topic and objective stated in the Table of Specifications, the notation used in the item was consistent with that used in the written course material, and the options were correct; second, based on experts' judgment, the item consisted of a clear statement; third, the item difficulty was in a range of 0.10 to 0.90 and the item discrimination equal to or greater than 0.30. An item was revised and called a revised item: first, if the item consisted of incorrect options, inconsistent notation, and umbiguous statement; second, if the item difficulty was less than 0.10 or greater than 0.90, and/or the item discrimination was less than 0.30 but it was not a negative value. However, in practice, if the item difficulty was less than 0.10 or greater than 0.90 but the item discrimination was equal to or greater than 0.30, the reviser decided whether to revise the item or not by

considering the correctness of the content. In the same way, if the item discrimination was less than 0.30 but the item difficulty was in a range of 0.10 to 0.90, the reviser decided whether or not to revise the item by considering the correctness of the content. In practice, some revisers also rejected the revision of an item which had unacceptable value of both item difficulty and discrimination because they considered that the item did not contain an error and the statement was also clear. An item was rejected if, first, the item contained a fatal error which could not be revised such as inconsistency between the item and the Table of Specifications, and second, the item difficulty was less than 0.10 or greater than 0.90, and the item discrimination was less than 0.10 (Zainul, 1992).

The number of revised, non-revised, and rejected items produced from revision activity in each faculty is not the same. In the Faculty of Education (FKIP), for example, the revision activity produced 32,481 revised items (about 60%) and about 40% non-revised items. In the Faculty of Mathematics and Natural Science (FMIPA), the revision activity produced 4,753 revised items (about 60%) (Source: UT's Exam-Processing Center). UT's Exam-Processing Center faced problems in obtaining information about the percentage of rejected items in

both faculties. To date, most of the revised and non-revised items have not been re-used on new examinations.

Purposes and Definitions

Purposes. The purpose of conducting this study was to investigate the differences in the quality of UT's test items in terms of item difficulty, discrimination, and test reliability. A comparison was made between old tests and new tests, between original items and revised items, and between non-revised items in old and new tests. The new test was composed of revised and non-revised items of old tests. Items and tests were compared across faculties as well. An investigation was made of the extent to which the test items used at both faculties met the two sets of criteria proposed by Ebel (as the expert) and by UT's policy. Ebel's criteria was used in this study because UT's criteria is based upon Ebel's. UT is a new institution and it decided to use a more generous set of criteria than Ebel's criteria.

Definitions. For norm-referenced test developers who have a goal to maximize test-score variance and reliability, good tests have been defined as reliable tests having sufficient items with appropriate discrimination and difficulty. A good test should have over 25% of items with discrimination values of 0.40 and up, less than 25% of items with discrimination values in

a range of 0.20 to 0.39, less than 15% of items with discrimination values from 0.01 to 0.19, and less than 5% with zero or negative discrimination (Ebel, 1972). A good test should also have item difficulties in the range from 0.30 to 0.70, averaging about 0.50; also it should have a high reliability (greater than or equal to 0.80) (Allen & Yen, 1979; Ebel, 1972). Allen and Yen (1979) argued that generally, item difficulties that approximate the above standards maximize the information the test provides about differences among examinees. Since, maximum discrimination among examinees over all levels of performance is desired at Universitas Terbuka (UT), a range of difficulty from about 0.30 to 0.70 with an approximate average 0.50 was used in this research.

The discrimination index of an item was defined as a coefficient which indicates how well the item discriminates the good students from the poor students on the basis of total test score (Crocker & Algina, 1986; Ebel, 1972). The discrimination index used in this thesis was the Point Biserial Correlation Coefficient. The discrimination index has values in the range of -1 to 1, with the high positive values being the most desirable. An item with high positive value of discrimination index is desirable because the item is answered correctly by many high-scoring examinees and is

missed by low-scoring examinees.

Item difficulty, p, was defined as the proportion of students who answer correctly (Adkins, 1974; Crocker & Algina, 1986; Ebel, 1972). The item difficulty may take values in the range from 0 to 1. A p-value of 0 indicates that none of students answered the item correctly, while p=1 indicates that all students answered the item correctly.

The reliability coefficient used in this thesis was the internal consistency measure of Kuder-Richardson (KR 20). Internal consistency indicates how consistently the examinees performed across items or subsets of items on a single test form (Crocker & Algina, 1986).

General Overview

In test construction, a general goal is to arrive at a test of minimum length that will yield scores with the necessary degree of reliability and validity for the intended uses. This is typically accomplished by field-testing a large pool of items and selecting a subset of items from that pool that make the greatest contributions to reliability or validity. In constructing a new test or reconstructing an existing one, the final set of items usually is identified through a process known as item analysis. Item analysis is a term broadly used to define the computation of any

statistical property of examinees' responses to an individual test item.

There are several different item-analysis techniques proposed by test experts. Some item analyses use item discrimination and difficulty indices, while others use item reliability indices.

The item discrimination index is used whenever the purpose of the test is to provide information about individual differences on the construct measured by the test. Norm-referenced test developers, for example, who consider a student's performance as a comparison of his or her score with the average score of his or her peers (class), greatly emphasize the discriminating ability of items. Therefore, they use the item discrimination index in measuring that ability. In this case, the item discrimination index is an index of how effectively the item discriminates between examinees who are relatively high on the total test score and those who are low. The goal is to identify items for which high-scoring examinees have a high probability of answering correctly and low-scoring examinees have a low probability of answering correctly. Such an item discriminates between examinees who know the material and those who do not. In contrast, an item on which both high- and low-scoring examinees have the same probability of answering

correctly does not discriminate between good and poor students. It would be even less desirable to have items that are missed by many high-scoring examinees but are answered correctly by low-scoring examinees. Such items are said to show negative discrimination.

There are no ideal values of item discrimination. Ebel (1972) and Lien (1976) proposed different evaluation criteria for the values of the discrimination as shown in Table 2.

Table 2
Evaluation of Item Discrimination

Ebel		Lien	
Index of discrimination	Item evaluation	Index of discrimination	Item evaluation
0.40 and higher	Very good items	Over 0.50	Good Items
0.30 to 0.39	Reasonably good, but possibly subject to improvement		
0.20 to 0.29	Marginal items, with considerable need for improvement	0.20 to 0.50	Fair items
Below 0.20	Poor items which should be rejected or improved by revision	Below 0.20	Poor items

From Table 2, it can be seen that Lien's criteria is slightly more rigorous than Ebel's criteria. In terms

of all items in a test, Lien argued that 50% of the items should have discrimination exceeding 0.40, less than 40% should have values between 0.20 and 0.40, less than 10% should have values between 0 and 0.20, and none have negative values. Ebel, on the other hand, suggested that over 25% of the items should have discrimination 0.40 and up, less than 25% should have values in a range of 0.20 to 0.39, less than 15% should have values in a range of 0.01 to 0.19, and less than 5% have zero and negative values. Ebel and Lien proposed these criteria based on their experiences with a wide variety of classroom tests. However, these criteria were used only as a guide to interpretation. According to Ebel and Lien, although these criteria are reasonable for most classroom tests, the criteria may not all be appropriate to certain special tests. If an instructor feels that the criteria are not suitable, he or she may choose more appropriate criteria and re-evaluates the tests in terms of these new criteria.

Values of item difficulty are more problematic. The value of item difficulty does not solely indicate the property of the item, but it also reflects the ability of the group of students responding to that item. Hence, an item difficulty of p=0.56 means that when the item was administered to that particular group

of students, 56% correctly responded. The p-values of an item fluctuate across test administrations because of the different abilities of students. In addition, other variables, such as guessing and the conditions under which the test is given, also affect the value of item difficulty.

Disagreement exists among test experts about the appropriate range of item difficulties. Generally, test experts consider that a good norm-refenced test must include some easy items to test the low achievers and some difficult items to test the high achievers. They consider that the difficulty levels of test items should be greater than chance and in a range of 0.30 to 0.70 (Allen & Yen, 1979; Ebel, 1972). At Universitas Terbuka (UT), the testing system have tended toward norm-referenced measurement. Under this system, UT accept items with difficulties in a range of 0.20 to 0.90 and discriminations equal to or higher than 0.20.

To realize the purpose of this thesis, then, it is first necessary to describe and define the various item-analysis techniques and to investigate test experts' opinions about the criteria of item discrimination, item difficulty, and test reliability. A next step will involve an exploration of previous studies in which item characteristics have been studied.

Research Questions

Since only the Faculty of Mathematics and Natural Science (FMIPA) and the Faculty of Education (FKIP) have developed tests based on UT's standard procedure, a question of the quality of tests based on item difficulty, item discrimination, and test reliability in both faculties arises. What is the distribution of item difficulties, item discriminations, and test reliabilities within both faculties? Is there an effect of revisions on those item characteristics within both faculties?

In more specific terms, two research questions were posed:

1. Are there differences in the quality of item difficulty, item discrimination, and test reliability between faculties?
2. Do old tests and new tests consisting of both revised and non-revised items differ in the quality of item difficulty, item discrimination, and test reliability?

These questions are answered by completing a descriptive analysis. The purpose of this analysis was to examine the differences in qualities of item difficulty, item discrimination, and test reliability between old tests and new tests, between original and revised items, and

between non-revised items of old and new tests in the two faculties. An investigation was made of the extent to which the test items used at both faculties met the two sets criteria proposed by Ebel as the expert, and by UT's policy.

CHAPTER II

Review of the Literature

The literature review for this study was organized around three major themes: first, a discussion of three methods of item analysis; second, an investigation of experts' opinion about criteria of item discrimination, item difficulty, and test reliability; third, an exploration of previous studies in which these item characteristics and item analyses have been studied.

Methods of Item Analysis

Item analysis has been recognized as a descriptive procedure for determining the characteristics of test items. For norm-referenced test (NRT) developers who compare an individual examinee's performance with the performance of others, the characteristics generally described include item difficulty and item discrimination. However, for criterion-referenced test (CRT) developers who compare an individual examinee's performance with a content standard score regardless of how others had performed on the test, the characteristics generally described include item difficulty, item sensitivity to instruction, and agreement indices. The concept of item difficulty used in CRT is the same as that used in NRT but the criteria used in judging acceptable levels of item difficulty are

different. The concept of item sensitivity to instruction in CRT is analogous to the concept of item discrimination in NRT. While item discrimination has been used as a measure of how well that item discriminates between good and poor students, the item sensitivity to instruction has been used as a measure of how well that item discriminates between students who have received instruction and those who have not. The agreement indices have been used to examine degree of agreement among response patterns for particular items. For example, do these two items measure the same skill or content; what proportion of examinees passed or failed both items; and did examinees perform significantly better on one item than on the other (Crocker & Algina, 1986)?

Both NRT and CRT developers have used both classical true score (CTS) and item response theory (IRT) models in determining the characteristics of test items. Those who have used the CTS model deal with the test score $X = T + E$ where T represents an individual's true score and E is a random error. They have observed the pattern of test scores and observed examinees' performance on the test, but have not provided information about how examinees at different ability levels for a trait performed on an item. Those who have

used the IRT model deal with item score. They have observed the pattern of item responses and observed how examinees at different ability levels for a trait performed on an item (Crocker & Algina, 1986).

In the next section, three methods of item analyses are discussed. The first and second methods have made use of the CTS model; while the last method have followed the IRT model.

Item analysis using item difficulty and item discrimination indices. One of the most suitable item analysis techniques for classroom teachers or test developers who do not have access to a computer is item analysis using item difficulty and item discrimination index (Allen & Yen, 1979).

The item difficulty for item i, p_i, has been defined as the proportion of students who answer that item correctly (Adkins, 1974; Allen & Yen, 1979; Crocker & Algina, 1986; Ebel, 1972). This proportion is really inversely related to the common usage of the word difficulty, because the larger the proportion of students who answer correctly, the easier the item. From long usage, however, the higher the item difficulty the easier the item has been understood to be (Adkins, 1974). In using the item difficulty indices, the assumption that students have sufficient time to attempt

every item is required. In special circumstances, when speed of performance on intrinsically easy items is being measured, the indices of item difficulty would not be particularly useful (Adkins, 1974).

The item discrimination index has been defined as a coefficient that indicates how well the item discriminates good students from poor students (Adkins, 1974; Allen & Yen, 1979; Crocker & Algina, 1986; Ebel & Frisbie, 1986). The item discrimination index also indicates the degree to which responses to one item are related to responses to other items in the test. Assumptions required in using item discrimination indices are that the total test measures what it should measure and also that it measures a unitary or single factor. If it, in fact, measures two or more quite different factors, then this type of item analysis may yield results that are difficult to interpret (Adkins, 1974). For example, if a teacher has combined thirty items calling for quantitative ability and fifty items depending heavily on verbal ability into one test, then blind application of item analysis using the total score on the eighty items as criterion might well lead to the discard of the thirty quantitative items even if they function effectively in regards to achievement of quantitative outcomes.

Many different methods have been developed to determine discrimination indices, such as the index of discrimination (U-L index) and the point biserial correlation coefficient (Allen & Yen, 1979; Crocker & Algina, 1986).

The index of discrimination has been calculated by the formula D = Pu - Pl, where Pu is the proportion of students in the upper group who answer the item correctly and Pl is the proportion of students in the lower group who answer the item correctly (Crocker & Algina, 1986). The upper and lower ranges generally are defined as the upper and lower 10% to 33% of the sample with students ordered on the basis of their total test scores (Allen & Yen, 1979). However, when the total test scores are normally distributed, using the upper and lower 27% produces the best estimate of D (Kelley, 1939). When sample size is reasonably large, virtually the same results can be obtained with the upper and lower 30% or 50% (Beuchert & Mendoza, 1979; Englehart, 1965).

The point biserial correlation coefficient has been defined as a correlation between item score and total test score. When performance on an item is uncorrelated with performance on all the other items in a test, the value of point biserial correlation still will be

positive, because the item score is included in the total test score (Allen & Yen, 1979). To control for this effect, an item point biserial can be calculated using test scores x', in which the item score is not included. For a 25-item test, for example, the point biserial for the second item would be based on the students' scores on the 24 items excluding item 2. However, if the number of items is reasonably large (25 or more), this fact is seldom a problem (Crocker & Algina, 1986).

Item analysis using item reliability index. When the goal of item selection is to maximize the internal consistency variability ($R_{xx'}$) or criterion-related validity (R_{xy}), the use of item analysis using item reliability indices is suggested (Allen & Yen, 1979).

To choose the best items, four statistics are required for each item i, that is, the item difficulty p_i, the item-score standard deviation $S_i = \sqrt{p_i\ (1 - p_i)}$, the item reliability index $S_i.R_{ix}^*$ where R_{ix}^* is the point biserial correlation between the item score and the N-item test score, and the item validity index $S_i.R_{iy}$ where R_{iy} is the item point biserial correlation between item score and the criterion score.

To choose items for a test with maximum internal consistency reliability, the item-score standard

deviation S_i is plotted on the vertical axis and the item reliability index $S_i.R_{ix}^*$ is plotted on the horizontal axis. The best items have high item score/test score correlation.

To choose items for a test with maximum validity, the item validity index $S_i.R_{iy}$ is plotted on the vertical axis and the item reliability index $S_i.R_{ix}^*$ is plotted on the horizontal axis. The best items would have validity indices close to their reliability indices.

Test developers should decide whether reliability or validity is the more important goal because items chosen to maximize validity may not produce a test having good internal-consistency reliability.

Item analysis using item-characteristic curves. An item-characteristic curve (ICC) has been defined as a graphical display of the relationship between the probability of passing a particular item and the students' position on the underlying trait that is measured by the test (Allen & Yen, 1979). Since scores on the underlying trait are generally not available, observed test scores are used as estimators of trait values. For each item, then, the ICC is estimated by a plot with total test scores on the horizontal axis and the proportion of students passing the item on the

vertical axis.

The rationale underlying item analysis using ICC is similar to the rationale for item analysis using item difficulty and discrimination indices. An ICC for a good item should display a positive slope and a moderate difficulty level. A positive slope has indicated positive discrimination indices. The item difficulty has been defined as total test score corresponding to a proportion of 0.50 on the vertical axis.

General consideration in the item analysis. In item analysis using difficulty and discrimination indices, the item difficulty can be easily altered by changes in the sample. On the other hand, the ICC is less easily affected by changes in the sample, and so item difficulty and discrimination obtained from an ICC will tend to be stable over samples of students (Allen & Yen, 1979). A requirement for an effective item analysis is that the test developer obtain appropriate samples and sample sizes for item tryout. However, there is no absolute rule for the minimum number of students to be used in an item analysis study (Crocker & Algina, 1986). As a general rule, most item parameters in an item analysis can be estimated with relative stability for samples of 200 students, so this might be considered the minimum number required. Another long

standing rule-of-thumb (Nunnaly, 1967) is to have 5 to 10 times as many subjects as items. However, Tinary (1979) considered 25 students to be a reasonable sample.

Summary. In sum, there are three different item-analysis techniques: those using item difficulty and item discrimination indices, those using item reliability and validity indices, and those using item characteristics curves (ICC). Among these techniques, only the ICC technique uses a concept of item response theory (IRT). This technique is less affected by changes in sample size and in ability levels of students because it observes the performance of students who have different ability level on an item. On the other hand, item analysis using the concept of classical true score (CTS), such as item analysis using item difficulty and discrimination indices, is easily affected by changes in sample size and in ability levels of students because it does not observe the performance of students who have different ability level. Since the ICC technique observes the performance of students who have different ability level on every item, this technique is more complex and difficult to apply at an institution which has a large number of items such as the Indonesian Open University (UT). For practical reasons, UT decided to use item analysis using item difficulty and

discrimination indices. ICC was not a possible choice to analyze items because UT is not prepared to complete ICCs at this time. Item analysis using reliability Indices was not used because the goal of item selection in UT is not to maximize the internal consistency variability or criterion-related validity, but to examine the discrimination and difficulty of each item. Therefore, for the purpose of this study, item analysis using item difficulty and item discrimination indices was used. The point biserial correlation coefficient was used to compute item discrimination values. This technique was expected to be relatively unaffected by changes in the sample size because a sample of courses having at least 30 students was selected.

Criteria of Item Discrimination, Item Difficulty, and Test Reliability

Criteria of item discrimination. Not all items need to discriminate between good students and poor students. Items on mastery or criterion-referenced tests do not need to discriminate between students, because the goal is for all students to respond correctly. In addition, the writer of criterion-referenced items is interested solely in constructing items to measure the objective in the most direct way.

Items that do not discriminate between more and less knowledgeble or skillful students do not need to be eliminated from a criterion-referenced test if they reflect important learning outcomes. However, as long as differences in student learning exist and as long as a major purpose of university testing is to identify such differences, the discrimination of each item should be examined (Ebel & Frisbie, 1986; Hopkins & Stanley, 1981). In an educational achievement test, where the principal function is to distinguish different levels of achievement as clearly as possible, it is desirable for every item to have as high a discrimination index as possible (Ebel, 1972). Adkins (1974) argued that the validity of an item is essentially the discriminating power of the right answer. If a large proportion of the good students answer an item correctly and a small proportion of the poor students answer it correctly, that item discriminates properly and contributes to the test purpose (Ebel & Frisbie, 1986). An item that does not discriminate between good and poor students is wasted when discrimination is a test purpose (Green, 1981).

A well-written test item will distinguish between students who know the answer and those who do not because of incomplete information, misinterpretation, or

misperception. In effect, a question's discriminating power depends on the reasoning sophistication required to select the correct answer from several alternatives which all seem plausible to students who are not thoroughly knowledgeable. To obtain high discrimination, the author of multiple-choice questions must write the stem and alternatives so that those who know what is being asked can answer correctly and those who are less knowledgeable
will be defeated by their ignorance. An ambiguous question will be ambiguous to poor and good students alike and thus will not discriminate between students with different levels of achievement.

According to Ebel and Frisbie (1986) and Lien (1976), experience with a wide spectrum of classroom tests suggests that the indices of item discrimination for test items can be evaluated in the terms described in Table 2 (page 19). As described in Table 2, Lien's criteria are slightly more rigorous than are Ebel's. In terms of all items in a test, Lien argued that 50% of the items should have discrimination exceeding 0.40, less than 40% should have values between 0.20 and 0.40, less than 10% should have values between 0 and 0.20, and none have negative values; while Ebel suggested over 25% of the items should have discrimination 0.40 and up,

less than 25% should have values in a range of 0.20 to 0.39, less than 15% should have values in a range of 0.01 to 0.19, and less than 5% have zero and negative values.

In comparing two alike tests, Ebel and Frisbie (1986) argued that the test in which the average index of item discrimination is the higher will always be the better test, that is, the more reliable.

Criteria of item difficulty. There is no hard and fast rule concerning optimum difficulty values for items to be included in a test. The choice of appropriate item difficulty depends on the purpose of the test. A test used to select graduate students for a university that admits about ten percent of the applicants should contain extremely difficult items. On the other hand, a test used to select children for a remedial education program should contain, in relative terms, very easy items (Allen & Yen, 1979). Tests for criterion-referenced measurement such as mastery tests, minimum-competency tests, and some professional certification tests should not contain items that are moderate in difficulty because criterion-referenced test (CRT) developers have not been concerned with selecting items to maximize test-score variance and reliability (Ebel and Frisbie, 1991), but they have been concerned with

selecting items to maximize the degree of matching between an item and a learning objective (Payne, 1974). In a mastery test, for example, if every student masters the material and if all test items are tied closely to specific objectives, nearly every student should do well on all items. Therefore, it may be reasonable to expect and select items at an item difficulty of about 0.80, if instruction has been effective. However, if the primary goal of an instructor is to maximize test-score variance and reliability, as should probably be the case for most classroom tests that use norm-referenced tests in order to identify the individual differences in achievement, moderately difficult items with high discrimination should be chosen (Ebel & Frisbie, 1986, 1991).

An item offers the maximum amount of information about differences among students when the difficulty is 0.50. This would suggest that all items should have a difficulty of 0.50, but the usefulness of this suggestion is influenced by intercorrelations among items. In the extreme, if all items intercorrelated perfectly and had difficulties of 0.50, half of the examinees would receive a total test score of 0, and the other half would receive a perfect total score. Therefore, there would be no fine discrimination among students' ability levels. Therefore, it is best to

choose items with a range of difficulties that average 0.50 (Allen & Yen, 1979).

Disagreement has existed among test experts about the range of item difficulties for norm-referenced tests. Some experts have considered that a range of difficulties about 0.30 to 0.70 is the best (Allen & Yen, 1979; Lien, 1976). According to Lien, and Allen and Yen, most of the items should be in that range so that the majority of the students can have a reasonable opportunity to answer the questions. On the other hand, an item with an item difficulty of less than 0.30 is too difficult and could reduce student motivation. A few items should be easy to encourage the students to attempt the test items and a few should be difficult in order to challenge the better students. A slightly different range was proposed by Hopkins and Stanley (1981) who considered a range of difficulties from 0.25 to 0.75 as to be the best range because all items in that range have potential for high discrimination (0.50 and up). Wood (1961), considered that the items should vary in difficulty through a range of 0.15 to 0.85. Henryssen (1971) determined the range of difficulties based on the homogeneous tests. A test is said to have item homogeneity and is called a homogeneous test when examinees perform consistently across items within the

test. In order for a group of items to be homogeneous, they must measure the same type of performance (or represent the same content domain). The items must also be well written and free of technical flaws that may cause examinees to respond on some basis unrelated to the content (Crocker & Algina, 1986). According to Henryssen, the range of item difficulty should be rather narrow for heterogeneous tests and wider for homogeneous tests. Henryssen suggested a range of item difficulty from 0.40 to 0.60 if the average correlation between the item scores and the total test score is between 0.30 and 0.40. If the average correlation of item score/total test score is higher than 0.40 (that is, if the test is more homogeneous), a wider range of item difficulties is suggested. If the average correlation of item score/total test score is less than 0.30, a narrower range of item difficulties can be used.

Criteria of test reliability. For norm-referenced test (NRT) developers, the reliability of the scores for a group of students is the most important statistical measure of the quality of the scores. The goals of high score variability, high discrimination, and moderate difficulty, all contribute to the major aim of achieving high test-score reliability. Hopkins and Stanley (1981) suggested setting standards for reliability of at least

0.90 for standardized tests such as those used for college admission. However, for classroom tests, the standards for reliability are generally lower than those of standardized tests. Ebel (1972) suggested setting standards for reliability of 0.80 or more for classroom tests.

By contrast, the writers of criterion-referenced tests (CRT) have considered the value of standard error measurement, which indicates the degree of measurement precision, more meaningfully than the value of classical reliability coefficient used in NRT because the classical reliability coefficient used in NRT depends upon the existence of differences or variabilities among students' test scores (Hopkins & Stanley, 1981; Mehrens & Lehmann, 1973). Since CRT are used most often in mastery testing situations, an individual who attempts an item representing an objective and has prepared conscientiously for the test is highly likely to answer the item correctly. His or her total score is likely to be as high as are the scores of the total group. Therefore, variability among scores is likely to be low, and so, the classical reliability coefficient will tend to be low, that is, it will underestimate reliability. Some experts proposed alternative approaches for estimating reliability of CRT, such as P, P*, Cohen's

kappa, Livingston's K2(X,T), and Brennan and Kane's M(C). For CRT, Payne (1974) suggested the reduction of standards of acceptable reliability to about 0.50 to accommodate the lowered variability.

Summary. Several different criteria for examining item discrimination, item difficulty, and test reliability have been reviewed.

For criterion-referenced test developers, the goal is to make sure an item is an accurate reflection of a learning objective. Difficult or easy, discriminating or indiscriminating, the important thing is to make the item represent the class of behaviors stated in the objective. In terms of reliability, Payne (1974) suggested to reduce standards of acceptable reliability to about 0.50 to accomodate the lowered variability.

For norm-referenced test developers, the goal is to maximize variant scores. In order to produce maximum score variability, they disdain items which are "too easy" or "too hard" and select items with moderate difficulty and high discrimination. However, there exists disagreement among norm-referenced test experts about the criteria for discrimination values and the range of difficulty levels. Lien (1976) proposed a slightly more rigorous criteria of discrimination values than those advanced by Ebel (1972). Some experts

proposed a range of difficulties about 0.30 to 0.70, while the others advanced a range of 0.25 to 0.75 or 0.15 to 0.85. Another expert determined ranges of difficulty based on the homogeneous tests. For test reliability, Ebel suggested setting standards for reliability at 0.80 or more.

For the purpose of this study, the criteria for norm-referenced tests were used. Item difficulty in a range of 0.30 to 0.70 and Ebel's criteria for evaluating item discrimination and test reliability were used. Ebel's more generous criteria were used because UT is a new institution. As banks of test items are enlarged and improved, the criteria can become more rigorous.

Exploration of Previous Studies

Little attention has been devoted to studying the quality of test items based on the distribution of item difficulty and item discrimination. One notable exception is a study by McMillan, Mundrake, and McGuire (1989). They evaluated the quality of tests at a business school based on the distribution of item difficulty and item discrimination. In determining the quality of the test, they used frequency distribution and summary statistics for item difficulty and item discrimination. They compared the difficulty and

discrimination statistics of instructor-prepared multiple-choice tests to the recommendations of testing experts and found that a large percentage of the tests failed to meet the discrimination and difficulty standards proposed by several testing scholars.

On the other hand, much attention has been devoted in the literature to studying the effects of item order and item format on item difficulty, item discrimination, and test reliability. Although major texts on classroom measurement advocate an easy-to-hard ordering of items (Gronlund, 1985; Mehrens & Lehmann, 1984), research on such topics is inconclusive. For example, some studies (Balch, 1989; Bresnock, 1989; Hambleton & Traub, 1974; Plake et al., 1982) have reported significant differences in students' total test scores when variations in item order by statistical difficulty were studied. Others (Allison, 1984; Huck & Bowers, 1972; Monk & Stallings, 1970; Plake, 1980; Sander, 1988) have reported no differences. Most of those investigators used experimental design using an easy-to-hard ordering of items for the first group, hard-to-easy ordering for the second group, and random ordering for the third group.

Most investigators studying the effects of item format on item characteristics used one correct response

option and inclusive alternatives "None of the Above" as the correct response and as the distractor. Methodology has varied considerably across "None of the Above" studies in the literature. Some used experimental design using regular items for one group and for another group used the same items modified by dropping an existing option and adding "None of the Above" (Dudycha & Carpenter, 1973; Kolstad, 1991; Tollefson, 1987). Some used a naturalistic setting over a variety of subject areas (Frary, 1991). Furthermore, contradiction exists among research results about the different effects of complex response alternatives such as "None of the Above" and correct response options on item difficulty, item discrimination, and test reliability. Some researchers found that there was no effect on item difficulty, discrimination, and test reliability associated with the "None of the Above" alternative on a test administered to students (Wesman & Bennett, 1946), whereas other researchers found the items with the "None of the Above" option were more or slightly more difficult than the items with single correct response options (Dudycha & Carpenter, 1973; Frary, 1991; Kolstad, 1991; Mueller, 1975; Oosterhof & Coats, 1984; Tollefson, 1987).

The effect of item revision on item discrimination

is another area which has largely been ignored in the literature. Lange, Lehmann, Mehrens (1967) conducted one such study using a test consisting of non-revised items, revised items, and new items. In analyzing their data, they used frequency distribution and summary statistics for item discrimination. They found that revised items showed an improvement in discrimination. Some of the non-revised items improved in discrimination, while others decreased. They also found that the average discrimination of non-revised items was lower than those in the new test, while the average of discrimination of the revised items was higher than those of the original items. The revised items also had a higher discrimination than the new items written specifically for the exam.

Item revision usually is conducted in the development and validation of a test. The test developer administers the test to a sample of students and evaluates the item analysis using item difficulty, discrimination, and test reliability indices. Those items which are not functioning well are reviewed and considered for possible revision. These items are administered to a second independent sample of students, and then the item discrimination, difficulty, and test reliability of their scores are re-evaluated. Padilla,

Cronin, and Twiest (1985) performed this procedure in their study of the development and validation of a test of Basic Process Skills. From the first field trial, they found that point biserial correlations (discrimination indices) for 30 of the 36 items were above 0.20 with the average of 0.34, that item difficulties ranged from 0.20 to 0.90 with an average value of 0.62, and that the test reliability (KR 20) was 0.78. These results indicated that most items were functioning well. However, Padilla, Cronin, and Twiest reviewed and revised those items with point biserial correlation below 0.30. From the second field trial, they found that point biserial correlations for 33 of the 36 items were above 0.25 with an average of 0.38, that item difficulties ranged from 0.27 to 0.95 with an average of 0.75, and that the test reliability was 0.82. Since they considered that the item discrimination and difficulty indices fell within the acceptable range for reliable tests proposed by Payne (1974), they decided that no further test revision was necessary.

Summary. Only a few studies have been devoted to the study of item quality and item revision based on distribution of item difficulty, discrimination, and test reliability indices, while much attention has been devoted to studying the effects of item order and item

format on these indices.

Those who have evaluated the quality of test items are McMillan, Mundrake, and McGuire (1989). They compared the difficulty and discrimination of the instructor-prepared multiple-choice test items to the recommendation of testing experts by using descriptive analysis. Lange, Lehmann, and Mehrens (1967) also used descriptive analysis in analyzing the effects of item revision on item discrimination. They used a test consisting of non-revised items, revised items, and new items.

Those who have studied the effects of item order on item difficulty, discrimination, and test reliability indices are Allison (1984), Balch (1989), Bresnock (1989), Hambleton and Traub (1974), Huck and Bowers (1972), Monk and Stallings (1970), Plake (1980), Plake et al. (1982), and Sander (1988). Other researchers (Dudycha & Carpenter, 1973; Frary, 1991; Kolstad, 1991; Mueller, 1975; Oosterhof & Coats, 1984; Tollefson, 1987; Wesman & Bennett, 1946) have investigated the effects of item format on those indices. Research on such topics is inconclusive. Some studies reported significant differences in students' total test scores when variations in item order by statistical difficulty were studied, while others have reported no differences.

Some researchers also found that there were no effects on item difficulty, discrimination, and test reliability associated with the "None of the Above" alternative on a test administered to students, whereas other researchers found the items with the "None of the Above" option were more or slightly more difficult than the items with single correct response options.

CHAPTER III

Methodology

Introduction

In 1989, Universitas Terbuka (UT) revised all of the items which were offered by test administration 89.2. In revising the test items, UT used both judgmental and empirical methods with the use of empirical methods preceding the application of judgmental strategy. In using the empirical method, UT employed the statistical data of students' responses from the actual examinations, such as the item difficulty and discrimination. UT asked the revisers to identify the poor items based on the value of item difficulty and discrimination. Before revising the selected items, UT also asked the revisers to judge the clarity of the items, the correctness of the options, the appropriateness of the difficulty level, the consistency of the items and Table of Specifications, and the consistency of notations used in the tests and in written course materials. The items which were considered to be poor based on the two methods or based on the revisers' judgments only were revised. Those items which were considered to be good were used as first written. Until now, UT has not yet analyzed the statistical data of students' responses on the revised

and non-revised items in the new tests.

The methodology of this study was proposed in an attempt to answer the following research questions:

1. Are there differences in the quality of item difficulty, item discrimination, and test reliability between faculties?
2. Do old tests and new tests consisting of both revised and non-revised items differ in the quality of item difficulty, item discrimination, and test reliability?

Approach

The basic approach of this study was to describe the item characteristics of a sample of courses from three test administrations, on the basis of the analysis of students' responses on the revised and non-revised items in the old and new tests and then to evaluate the effects of the revisions on item characteristics.

Sample and Procedure

Sample. This study involved test results of six courses from each of two faculties at UT from three tests administrations. Initially, test administrations 87.1, 89.1, and 92.1 were used as samples in the two faculties, the Faculty of Mathematics and Natural Science (FMIPA) and the Faculty of Education (FKIP). However, since the data file for test administration

87.1 was not available at FMIPA, test administration 88.1, 89.1, and 92.1 were used as samples. FKIP still used test administration 87.1, 89.1, and 92.1 as samples.

The test in 92.1 was assembled of revised and non-revised items from tests 87.1 and 89.1 for FKIP, and of revised and non-revised items from tests 88.1 and 89.1 for FMIPA. The revised items are ones which were revised for two reasons. First, the items consisted of incorrect options, inconsistent notations, and ambiguous statements. Second, the items were either too difficult (the item difficulties were less than 0.10) or too easy (the item difficulties were greater than 0.90) and/or had low discriminations (the item discriminations were less than 0.30). The non-revised items are the items which were not revised because they were consistent with the topics and objectives stated in the Table of Specifications, had clear statements, and had appropriate difficulties and discriminations.

The tests were developed from the parallel Table of Specifications. This means that a test in 88.1, 89.1, and 92.1 for FMIPA and a test in 87.1, 89.1, and 92.1 for FKIP employed the same Table of Specifications, but the number of items developed from the Table of Specifications might differ from what was stated in the

Table. The number of items developed for a specific course might also differ across test administrations. For example, the number of items for Linear Algebra I in test administration 88.1 is 25 and in 89.1 is 30, while the number of items stated in the Table of Specifications is 30. For Statistical Method I, as another example, the number of items in 88.1 and 89.1 is 30, while the number of items stated in the Table of Specifications is 34. In this study, the tests in 92.1 consist of 30 items for FMIPA and 50 to 70 items for FKIP. The length of those tests were determined by the assembler in both faculties.

A purposive sampling technique of courses within faculty was conducted. This technique is a kind of non-probability or non-random sampling technique where the courses to be included in the sample were selected on the basis of judgment of their typicality: the type of the courses, the type of items, the time of administration of their tests, and the number of students who took the courses. In this way, from each of two faculties (FMIPA and FKIP), two programs were chosen and from each program, three courses were chosen as samples. All programs in FMIPA, both Mathematics and Statistics programs, were used in this study. One of four science programs and one of two social programs in

FKIP were used. Physics and English programs were selected as samples. From each program in each faculty, three courses were selected. The courses were selected based on four criteria. First, the courses were not general courses (the courses which should be taken by all students without considering their programs) such as the Indonesian Philosophy 'Pancasila' and Basic Natural Science. Second, the tests of the courses were offered both in 87.1 and 89.1 for FKIP, and in 88.1 and 89.1 for FMIPA. Third, the tests were offered in multiple-choice forms. Fourth, at least 30 students took each course.

Based on the criteria described above, Calculus I (MATK4110), Advanced Calculus I (MATK4212), and Linear Algebra I (MATK4112) were selected as samples from the Mathematics program and Statistical Method I (STAT4110), Survey Sampling Method (STAT4334), and Applied Experimental Design (STAT4213) were selected as samples from the Statistics program in FMIPA. In FKIP, Alternating Current (PFIS4439), Atomic Physics (PFIS4438), and Introduction to Quantum Mechanics (PFIS4437) were selected as samples from the Physics program. English for Arts and Science (PING4447), English for Education (PING4441), and Business English (PING4448) were selected as samples from the English program.

Procedure. UT's computer system provides examination scanning, statistical analysis, and grade reporting. The system was designed to process tests and maintain a log of each test administration that included the test administration code, the course code, the faculty code, the number of items, and the number of students. This summary-log information was used to help in isolating the test data to be analyzed in this study. This study used final exams which met four criteria: (a) they were not general courses (the courses which should be taken by all students without considering their programs); (b) the items were entirely multiple-choice; (c) they were offered both in 87.1 and 89.1 for FKIP, and in 88.1 and 89.1 for FMIPA; (d) the tests were taken by at least 30 students. With these restrictions, three courses were selected from each program in each faculty. For the selected courses, the revised and non-revised items from tests 87.1 and 89.1 used in FKIP were reviewed and assembled into approximately half revised and half non-revised items for test administration 92.1. The revised and non-revised items from tests 88.1 and 89.1 used in FMIPA were also reviewed and assembled into approximately half revised and half non-revised items for test administration 92.1.

In assembling the items of Advanced Calculus I,

most of the items in the test 92.1 of Advanced Calculus I were not assembled from the items from test 88.1 or 89.1 (Only 7% of items were assembled from tests 88.1 and 89.1 respectively). The assembler also did not remember the test administrations from which he had assembled the items for test 92.1. Therefore, the course of Advanced Calculus I was not part of the sample.

In the assembling items for Calculus I (MATK4110), eight of thirty items in Calculus I were not assembled from test administration 88.1 or 89.1, but were taken from test 88.2. Three of fifty items of Atomic Physics (PFIS4438) were not assembled from test administration 87.1 or 89.1, but were drawn from test 87.2. Moreover, since the number of revised items of a course in tests 88.1 and 89.1 at FMIPA is far less than the number of non-revised items, the goal of composing test 92.1 into 50% revised and 50% non-revised items was not fulfilled. As a result, most of the courses at FMIPA were composed of about 20% revised items and 80% non-revised items. Only one course, Calculus I (MATK4110), was composed of 43% revised items and 57% non-revised items. A more clear insight about the composition of revised and non-revised items in the new tests from FKIP and FMIPA is presented in Table 3.

Table 3
Composition of Items in Test Administration 92.1

Courses	The percentage of items assembled from each test administration					The percentage of revised items	non-revised items
	87.1	87.2	88.1	88.2	89.1		
				FKIP			
Physics							
PFIS4439	42.86	-	-	-	57.14	53.06	46.94
PFIS4438	54.00	6.00	-	-	40.00	56.00	44.00
PFIS4437	51.02	-	-	-	48.98	44.90	55.10
English							
PING4447	50.00	-	-	-	50.00	41.43	58.57
PING4441	55.00	-	-	-	45.00	50.00	50.00
PING4448	35.00	-	-	-	65.00	41.67	58.33
				FMIPA			
Mathematics							
MATK4110	-	-	30.00	26.67	43.33	43.33	56.67
MATK4112	-	-	44.83	-	55.17	6.67	93.33
Statistics							
STAT4110	-	-	53.33	-	46.67	26.67	73.37
STAT4213	-	-	46.67	-	53.33	20.00	80.00
STAT4334	-	-	60.00	-	40.00	17.67	82.33

In addition to summary-log information, item analyses before deletion of some items for tests 87.1, 88.1, 89.1, and 92.1 were used. Those item analyses were conducted before some items were deleted in order to improve students' scores. Those item analyses were

used because they better represented the quality of the items than the item analyses after deletion of some items. However, for some tests, such as test 87.1 of English for Education (PING4441) and Atomic Physics (PFIS4438), test 88.1 of Linear Algebra I (MATK4112) and Statistical Method I (STAT4110), test 89.1 of Linear Algebra I (MATK4112) and Survey Sampling Method (STAT4334), and test 92.1 of Alternating Current (PFIS4439) and Introduction to Quantum Mechanics (PFIS4437), the item analyses before deletion of some items could not be collected for this study due to problems in administration. Therefore, for those tests, the item analyses after deletion of some items were used. It was recognized that eliminating some items from a set of items affected the item characteristics and could lead to spurious results, especially on the value of item discrimination (Cureton, 1950). However, since the percentage of eliminated items was relatively small (less than or equal to 10%), the effect of item elimination on item discrimination was slight. In FKIP, only 2% of items was deleted from test 87.1 of Atomic Physics (PFIS4438) and 4% of items was deleted from test 87.1 of English for Education (PING4441). Respectively, 2% of items was deleted from test administration 92.1 of Alternating Current (PFIS4439) and Introduction to

Quantum Mechanics (PFIS4437). In FMIPA, 3% of items was deleted from test administration 88.1 of Statistical Method I (STAT4110), 4% and 7% of items was deleted respectively from test 88.1 and 89.1 of Linear Algebra I (MATK4112), and 10% of items were deleted from test 89.1 of Survey Sampling Method (STAT4334). Frary (1991) also used the item analyses after deletion of some items in studying the effects of the "None of The Above" option on item difficulty and discrimination.

The item analysis printout at UT listed the test administration code, the course code, the number of students taking the examination, the difficulty and discrimination indices of each test item, the mean of test score, the standard deviation of scores, the test reliability coefficient, and the standard error of measurement.

The printout data for test items from tests 87.1, 88.1, 89.1, and 92.1 were entered into a database file for analysis with coding for faculty, program, number of test items, status of the items: revised or non-revised, the test administrations from which the items were assembled for test 92.1, and for item-analysis information such as course code, test administration code, the number of students, the item difficulty and discrimination, and the test reliability. Database

files were analyzed using an SPSS program.

A descriptive analysis was used in analyzing the differences in quality of item difficulty, discrimination, and test reliability between old tests and new tests and those qualities between two faculties. Differences in item-characteristic qualities of old and new tests within and between faculties were analyzed through program-by-program and course-by-course comparison. Descriptive analysis was also used in analyzing the effects of the revisions on item difficulty and discrimination. In order to investigate whether or not the differences between the average item difficulty and discrimination of original and revised items were significant, a paired t-test was used. For a specific course, the differences of original and revised items' difficulty and discrimination indices were analyzed and course-by- course comparison was conducted. Program-by-program and faculty-by-faculty comparison were also analyzed. The differences of non-revised items' difficulty and discrimination indices in old and new tests were analyzed as well.

CHAPTER IV

Results

Findings for this study were grouped into three categories: those related to item difficulty, those related to item discrimination, and those related to test reliability. These findings were presented in order to answer the following research questions: (1) Are there differences in the quality of item difficulty, item discrimination, and test reliability between faculties? (2) Do old and new tests consisting of both revised and non-revised items differ in the quality of item difficulty, item discrimination, and test reliability ?

Item Difficulty

In order to judge whether item difficulties in both faculties, the Faculty of Mathematics and Natural Science (FMIPA) and the Faculty of Education (FKIP), had changed; and to determine whether the quality of item difficulty between faculties was different, two steps were taken. First, both Ebel's and UT's criteria for item difficulty were applied to the old (87.1, 88.1, and 89.1) and new (92.1) tests to determine if the distribution of item difficulties had changed. Differences in the mean of item difficulty and in the

percentage of items which meet the criteria for item difficulty were analyzed, and faculty-by-faculty, program-by-program, and course-by-course comparison were conducted. Second, both Ebel's and UT's criteria for item difficulty were applied to the revised and non-revised items on both old and new tests to determine the effects of item revision upon item difficulty. The effects of item revision upon item difficulty were analyzed through the analysis of the differences in the average item difficulty and in the percentage of items which have acceptable difficulty. Faculty-by-faculty, program-by-program, and course-by-course comparisons were conducted.

Comparison of old and new tests. As shown in Table 4, using either Ebel's or UT's criteria, the percentages of items within acceptable range of item difficulty in the Faculty of Education (FKIP) and the Faculty of Mathematics and Natural Science (FMIPA) were almost the same one test administration to the next. Under Ebel's criteria, only 50 to 60% items across the three test administrations in FKIP and FMIPA have acceptable difficulties. However, under UT's criteria which offers a wider range of item difficulty than Ebel's, 80 to 90% items have acceptable difficulties. This means that most items in FKIP and FMIPA met UT's criteria in item

Table 4
Distribution for Item Difficulty between Faculties

	Percent of items in each category					
	FKIP			FMIPA		
p-value	87.1	89.1	92.1	88.1	89.1	92.1
	Ebel's criteria					
Acceptable range 0.30-0.70	54.8	57.4	51.8	52.9	51.0	54.7
< 0.30	22.3	17.9	20.7	42.8	44.1	38.7
> 0.70	22.9	24.7	27.5	4.3	4.8	6.7
	UT's criteria					
Acceptable range 0.20-0.90	83.6	86.8	84.6	88.4	80.0	80.7
< 0.30	22.3	17.9	20.7	42.8	44.1	38.7
0.30-0.40	15.5	14.7	10.9	27.5	25.5	22.0
0.41-0.84	53.6	58.9	55.3	29.0	30.3	38.0
0.85-0.90	4.2	3.4	5.6	0.7	0.0	1.3
> 0.90	4.5	4.7	7.4	0.0	0.0	0.0

Note. Items with $p < 0.30$ are very difficult, p in 0.30-0.40 are difficult, p in 0.41-0.84 are moderate, p in 0.85-0.90 are easy, and $p > 0.90$ are very easy.

difficulty. This result is surprising because FKIP and FMIPA have not yet used the item difficulty indices listed in the item-analysis print-out as the basis for item selection or item assembly. In selecting or assembling the items, FKIP and FMIPA only select the items by matching the objectives and cognitive domains of the items with those stated in the Table of Specifications. It seems that the item selection or assembly procedure conducted in both faculties is likely to be the reason why 10-20% of items which do not meet the criteria were included in the tests. Even if UT used item analysis data in selecting test items, such items are often required to have an adequate and representative sampling of the course content.

Furthermore, in Table 4 it is shown that (1) the percentages of items in the Faculty of Mathematics and Natural Science (FMIPA) judged to be hard based on Ebel's ($p < 0.30$) or UT's criteria ($p \leq 0.40$) are higher than those in the Faculty of Education (FKIP), (2) the percentages of items in FMIPA judged to be moderate based on UT's criteria (p-values in a range 0.41 to 0.84) are lower than those in FKIP.

Under Ebel's criteria, the percentages of hard (p-values less than 0.30) and easy items (p-values greater than 0.70) in FKIP were almost the same (about 20%) while in

FMIPA, most of the items were hard (about 40%). Under UT's criteria, across the three test administrations, 30 to 40% items in FKIP were judged as hard ($p \leq 0.40$) while in FMIPA, they were about 70%. There were almost no easy items ($p \geq 0.85$) in FMIPA. 50 to 60% items in FKIP were judged as moderate (p-values in a range of 0.41 to 0.84) while in FMIPA, they were about 30%. These results were supported by data given in Table 5. The means of item difficulty in 13 of 15 FMIPA tests used in this study (86.7%) were less than 0.40, while 16 of 18 FKIP tests (88.9%) had high means (higher than 0.40) with 9 of the 16 tests being English courses. 2 of 9 tests (22.2%) in the FKIP Physics program had low means (less than 0.40) while no English tests had low means.

Furthermore, application of both Ebel's and UT's item difficulty criteria to the two old test administrations in comparison to the new test administration 92.1, leads to the conclusion that the distribution of item difficulties in the Faculty of Mathematics and Natural Science (FMIPA) and the Faculty of Education (FKIP) was relatively constant (see Table 4). Under Ebel's criteria, for example, the distribution of FMIPA and FKIP item difficulties in the old and new test administrations were similar.

Table 5
Distribution of the Proportion of Item Discrimination in a Test, Mean of Item Discrimination and Difficulty, and Test Reliability

Courses and period	Original N of items	N of items discarded	% of items retained	% of items in each range ≥ 0.40	0.20-0.39	0.01-0.19	≤ 0	Mean of discr	Mean of diff	N of student	KR-20
				FKIP							
PFIS4439											
87.1	40	0	100	7.7	56.4	30.8	5.1	0.246	0.422	148	0.586
89.1	50	0	100	12.8	46.8	40.4	0.0	0.249	0.517	62	0.702
92.1	50	1	98	14.0	60.5	14.0	11.6	0.261	0.480	108	0.727
PFIS4438											
87.1	45	1	98	9.8	41.5	39.0	9.8	0.199	0.346	152	0.427
89.1	60	0	100	14.3	37.5	39.3	8.9	0.223	0.424	62	0.696
92.1	50	0	100	2.1	48.9	40.4	8.5	0.184	0.416	166	0.464
PFIS4437											
87.1	45	0	100	19.5	41.5	26.8	12.2	0.244	0.402	60	0.656
89.1	50	0	100	2.2	44.4	46.7	6.7	0.195	0.349	107	0.511
92.1	50	1	98	8.3	60.4	27.1	4.2	0.247	0.400	147	0.685
PING4447											
87.1	70	0	100	27.5	49.3	17.4	5.8	0.287	0.527	196	0.812
89.1	80	0	100	14.7	48.0	32.0	5.3	0.247	0.650	94	0.810
92.1	70	0	100	11.9	53.7	28.4	6.0	0.254	0.618	95	0.796
PING4441											
87.1	70	3	96	19.7	42.4	28.8	9.1	0.242	0.553	108	0.772
89.1	80	0	100	34.7	46.7	17.3	1.3	0.320	0.525	156	0.894
92.1	60	0	100	35.2	44.4	16.7	3.7	0.313	0.589	143	0.849

table continues...

Courses and period	Original N of items	N of items discarded	% of items retained	% of items in each range ≥ 0.40	0.20-0.39	0.01-0.19	≤ 0	Mean of discr	Mean of diff	N of student	KR-20
PING4448											
87.1	70	0	100	23.9	49.3	19.4	7.5	0.275	0.619	239	0.855
89.1	60	0	100	39.0	42.4	10.2	8.5	0.325	0.593	123	0.863
92.1	60	0	100	21.8	47.3	21.8	9.1	0.272	0.657	91	0.802
				FMIPA							
MATK4110											
88.1	30	0	100	29.6	63.0	7.4	0.0	0.328	0.414	163	0.719
89.1	30	0	100	10.7	53.6	28.6	7.1	0.237	0.359	84	0.433
92.1	30	0	100	35.7	32.1	25.0	7.1	0.294	0.404	77	0.661
MATK4112											
88.1	25	1	96	54.5	31.8	13.6	0.0	0.368	0.378	67	0.750
89.1	30	2	93	22.2	59.3	18.5	0.0	0.312	0.353	87	0.667
92.1	30	0	100	14.3	42.9	35.7	7.1	0.237	0.333	105	0.475
STAT4110											
88.1	30	1	97	0.0	72.4	27.6	0.0	0.239	0.352	245	0.421
89.1	30	0	100	10.7	42.9	42.9	3.6	0.213	0.318	110	0.285
92.1	30	0	100	3.4	62.1	31.0	3.4	0.218	0.384	120	0.326
STAT4213											
88.1	30	0	100	0.0	65.5	31.0	3.4	0.227	0.319	194	0.389
89.1	30	0	100	14.3	35.7	39.3	10.7	0.207	0.330	57	0.265
92.1	30	0	100	17.2	41.4	20.7	20.7	0.224	0.386	29	0.420
STAT4334											
88.1	25	0	100	0.0	45.8	54.2	0.0	0.194	0.331	220	-0.056
89.1	30	3	90	26.9	46.2	26.9	0.0	0.297	0.371	91	0.642
92.1	30	0	100	44.4	14.8	29.6	11.1	0.290	0.394	29	0.692

In FMIPA, the distributions were positively skewed with the percentages of hard items (p-values less than 0.30) being far higher than those of easy items (p-values greater than 0.70). In FKIP, on the other hand, the distributions were nearly symetrical with the percentages of hard and easy items were almost the same. Although the distributions did not change very much, three points are noteworthy. First, the percentage of FMIPA items judged to be acceptable under Ebel's criteria was slightly higher in the new test administration 92.1 (52.9% in test 88.1, 51.0% in test 89.1, and 54.7% in test 92.1) while the percentage of FKIP items was not (54.8% in test 87.1, 57.4% in test 89.1, and 51.8% in test 92.1). However, the percentage of FMIPA revised items which were acceptable in the new tests 92.1 did not differ from the old tests 87.1 and 89.1 (the original items) while that of FKIP items was slightly higher (see Table 6).

Table 6
Percentage of Acceptable Items in FMIPA and FKIP Based on Ebel's Criteria

	Before revised (in the Old tests)	After revised (in the New tests)
FMIPA	47.1%	47.1%
FKIP	42.5%	46.9%

The percentage of FMIPA items judged to be moderate under UT's criteria was also higher in the new test administration 92.1 than the percentages in the old test administration 88.1 and 89.1, while the percentage of FKIP items was only slightly higher than the percentage in the old test administration 87.1 (see Table 4). Second, the percentage of FMIPA items judged to be hard under Ebel's ($p < 0.30$) and UT's criteria ($p \leqslant 0.40$) was slightly lower in the new test administration 92.1 than the percentages in the old test administrations 88.1 and 89.1 (Under Ebel's criteria, 38.7% of items in test 92.1 were hard while they were 42.8 % in test 88.1 and 44.1% in test 89.1. Under UT's criteria, 60.7% of items in test 92.1 were hard while they were 70.3% in test 87.1 and 69.6% in test 89.1). The percentage of FKIP items judged to be hard under UT's criteria was also slightly lower in the new test administration 92.1 than the percentages in the old test administrations 87.1 and 89.1 (37.8% in test 87.1, 32.6% in test 89.1, and 31.6% in test 92.1), but the percentages under Ebel's criteria were only slightly lower than the percentage in the old test administration 87.1 (22.3% in test 87.1, 17.9% in test 89.1, and 20.7% in test 92.1). Third, there were more items judged to be easy or very easy under both UT's ($p \geqslant 0.85$) and Ebel's criteria ($p > 0.70$) in the

FMIPA and FKIP new test administration. In FMIPA, under Ebel's criteria, 6.7% of items in test 92.1 were easy or very easy while they were 4.3% in test 88.1 and 4.8% in test 89.1. Under UT's criteria, 1.3% of items in test 92.1 were easy or very easy while they were 0.7% in test 88.1 and no easy items in test 89.1.

In FKIP, under Ebel's criteria, 27.5% of items in test 92.1 were easy or very easy while they were 22.9% in test 87.1 and 24.7% in test 89.1. Under UT's criteria, 13% of items in test 92.1 were easy or very easy while they were 8.7% in test 87.1 and 8.1% in test 89.1.

Differences in old and new tests concerning the quality of item difficulty in FKIP and FMIPA could also be seen from the distribution of individual item difficulty in their programs and courses (see Tables 7 and 8).

Within FKIP, considering both Ebel's and UT's criteria, the distribution of item difficulties in the new tests of the Physics and English programs did not change very much from those in the old tests (see Table 7). Under Ebel's criteria, for example, the distribution of item difficulties in the old and new tests of those programs were similar. In the Physics program, the distributions were positively skewed with the percentages of hard items (p-values less than 0.30)

Table 7
Distribution for Individual Item Difficulty within FKIP

p-value	Physics program	English program	PFIS 4439	PFIS 4438	PFIS 4437	PING 4447	PING 4441	PING 4448
	Percent of items in each category							
	Ebel's criteria							
87.1								
Acceptable								
range 0.30-0.70.	65.9	47.8	72.5	63.6	62.2	57.1	38.8	47.1
< 0.30	31.0	16.9	22.5	36.4	33.3	17.1	22.4	11.4
> 0.70	3.1	35.3	5.0	0.0	4.4	25.7	38.8	41.4
89.1								
Acceptable								
range 0.30-0.70.	60.0	55.5	56.0	66.7	56.0	37.5	71.3	58.3
< 0.30	27.5	10.9	16.0	25.0	42.0	13.8	11.3	6.7
> 0.70	12.5	33.6	28.0	8.3	2.0	48.8	17.5	35.0
92.1								
Acceptable								
range 0.30-0.70.	58.1	46.8	59.2	58.0	57.1	42.9	56.7	41.7
< 0.30	33.1	11.1	26.5	34.0	38.8	12.9	11.7	8.3
> 0.70	8.8	42.1	14.3	8.0	4.1	44.3	31.7	50.0

table continues...

p-value	Physics program	English program	PFIS 4439	PFIS 4438	PFIS 4437	PING 4447	PING 4441	PING 4448
	Percent of items in each category							
			UT's criteria					
87.1								
Acceptable								
range 0.20-0.90.	83.7	83.6	85.0	79.5	86.7	88.6	74.6	87.1
< 0.30	31.0	16.9	22.5	36.4	33.3	17.1	22.4	11.4
0.30-0.40	25.6	9.2	25.0	31.8	20.0	12.9	9.0	5.7
0.41-0.84	42.6	60.4	52.5	31.8	44.4	67.1	50.7	62.9
0.85-0.90	0.0	6.8	0.0	0.0	0.0	1.4	6.0	12.9
> 0.90	0.8	6.8	0.0	0.0	2.2	1.4	11.9	7.1
89.1								
Acceptable								
range 0.20-0.90.	85.6	87.7	88.0	85.0	84.0	80.0	93.8	90.0
< 0.30	27.5	10.9	16.0	25.0	42.0	13.8	11.3	6.7
0.30-0.40	24.4	7.7	26.0	21.7	26.0	2.5	12.5	8.3
0.41-0.84	45.6	68.6	52.0	51.7	32.0	61.3	72.5	73.3
0.85-0.90	0.6	5.5	2.0	0.0	0.0	6.3	3.8	8.3
> 0.90	1.9	6.8	4.0	1.7	0.0	16.3	0.0	3.3
92.1								
Acceptable								
range 0.20-0.90.	87.2	82.6	89.8	82.0	89.8	85.7	85.0	76.7
< 0.30	33.1	11.1	26.5	34.0	38.8	12.9	11.7	8.3
0.30-0.40	13.5	8.9	18.4	14.0	8.2	8.6	13.3	5.0
0.41-0.84	49.3	60.0	51.0	48.0	49.0	61.4	60.0	58.3
0.85-0.90	2.0	8.4	0.0	2.0	4.1	10.0	5.0	10.0
> 0.90	2.0	11.6	4.1	2.0	0.0	7.1	10.0	18.3

Note. Items with p < 0.30 are very difficult, p in 0.30-0.40 are difficult, p in 0.41-0.84 are moderate, p in 0.85-0.90 are easy, and those with p > 0.90 are very easy.

Table 8
Distribution for Individual Item Difficulty within FMIPA

p-value	Math program	Stat program	MATK 4110	MATK 4112	STAT 4110	STAT 4213	STAT 4334
			Percent of items in each category				
			Ebel's criteria				
88.1							
Acceptable							
range 0.30-0.70.	57.4	50.0	70.0	41.7	48.3	56.7	44.0
< 0.30	35.2	47.6	26.7	45.8	44.8	43.3	56.0
> 0.70	7.4	2.4	3.3	12.5	6.9	0.0	0.0
89.1							
Acceptable							
range 0.30-0.70.	60.3	44.8	66.7	53.6	53.3	46.7	33.3
< 0.30	37.9	48.3	33.3	42.9	43.3	50.0	51.9
> 0.70	1.7	6.9	0.0	3.6	3.3	3.3	14.8
92.1							
Acceptable							
range 0.30-0.70.	46.7	60.0	56.7	36.7	56.7	66.7	56.7
< 0.30	46.7	33.3	40.0	53.3	36.7	30.0	33.3
> 0.70	6.7	6.7	3.3	10.0	6.7	3.3	10.0

table continues...

p-value	Math program	Stat program	MATK 4110	MATK 4112	STAT 4110	STAT 4213	STAT 4334
		Percent of items in each category					
		UT's criteria					
88.1							
Acceptable							
range 0.20-0.90.	94.4	84.5	100	87.5	93.1	80.0	80.0
< 0.30	35.2	47.6	26.7	45.8	44.8	43.3	56.0
0.30-0.40	25.9	28.6	26.7	25.0	34.5	33.3	16.0
0.41-0.84	37.0	23.8	46.7	25.0	20.7	23.3	28.0
0.85-0.90	1.9	0.0	0.0	4.2	0.0	0.0	0.0
> 0.90	0.0	0.0	0.0	0.0	0.0	0.0	0.0
89.1							
Acceptable							
range 0.20-0.90.	81.0	79.0	83.3	78.6	90.0	76.7	70.4
< 0.30	37.9	48.3	33.3	42.9	43.3	50.0	51.9
0.30-0.40	31.0	21.8	36.7	25.0	40.0	20.0	3.7
0.41-0.84	31.0	29.9	30.0	32.1	16.7	30.0	44.4
0.85-0.90	0.0	0.0	0.0	0.0	0.0	0.0	0.0
> 0.90	0.0	0.0	0.0	0.0	0.0	0.0	0.0
92.1							
Acceptable							
range 0.20-0.90.	73.3	85.6	80.0	66.7	93.3	83.3	80.0
< 0.30	46.7	33.3	40.0	53.3	36.7	30.0	33.3
0.30-0.40	20.0	23.3	20.0	20.0	30.0	20.0	20.0
0.41-0.84	30.0	43.3	36.7	23.3	33.3	50.0	46.7
0.85-0.90	3.3	0.0	3.3	3.3	0.0	0.0	0.0
> 0.90	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Note. Items with $p < 0.30$ are very difficult, p in 0.30-0.40 are difficult, p in 0.41-0.84 are moderate, p in 0.85-0.90 are easy, and those with $p > 0.90$ are very easy.

being higher than the percentages of easy items (p-values greater than 0.70). In the English program, on the other hand, the distributions were negatively skewed with the percentages of easy items being far higher than the percentages of hard items. The distribution of item difficulties in the new tests of their courses also did not change very much from those in the old tests so as to have any impact on the measurement of student performance. Two points are important. First, applying Ebel's criteria to the two old tests (87.1 and 89.1) in comparison to the new test 92.1 leads to the findings that (1) none of the three Physics courses or the three English courses studied had a higher percentage of acceptable items in the new test than those judged to be acceptable in the two old tests, (2) a lower percentage of items judged to be hard ($p < 0.30$) in the new test than those judged to be hard in the two old tests only occurred in the course, English for Arts and Science (PING4447) (12.9% in test administration 92.1, 13.8% in test 89.1, and 17.1% in test 87.1). Second, applying UT's criteria to the two old tests in comparison to the new test leads to the finding that a higher percentage of acceptable items and moderate items, and a lower percentage of items judged to be hard ($p \leqslant 0.40$) only appeared in the course, Introduction to Quantum

Mechanics (PFIS4437). In this course, the percentages of acceptable items were 86.7% in test administration 87.1, 84% in test 89.1, and 89.8% in test 92.1. The percentages of moderate items were 44% in test administration 87.1, 32% in test 89.1, and 49% in test 92.1. The percentages of hard items were 53.3% in test administration 87.1, 68% in test 89.1, and 47% in test 92.1.

An interested finding also appeared in Business English (PING4448). In this course, the percentages of items judged to be acceptable under either Ebel's or UT's criteria were lower in the new test administrations 92.1 than the percentages in the two old test administrations 87.1 and 89.1. Under Ebel's criteria, the percentages of acceptable items in this course were 47.1% in test 87.1, 58.3% in test 89.1, and 41.7% in test 92.1; while under UT's criteria, the percentages of acceptable items were 87.1% in test 87.1, 90.0% in test 89.1, and 76.7% in test 92.1. The percentages of items judged to be moderate under UT's criteria were also lower in the new tests 92.1. In this course, the percentages of moderate items were 62.9% in test 87.1, 73.3% in test 89.1, and 58.3% in test 92.1. It seems that the decrease in the percentage of acceptable and moderate items in this course is not positively affected

by the evaluation criteria but by the nature of the items and characteristics of the examinees. Although test 92.1 consists of both the revised and non-revised items drawn from tests 87.1 and 89.1, which were examined by an expert, it can not be ensured that test 92.1 will have more acceptable items than tests 87.1 and 89.1. The item format and arrangement of items in the tests may affect the item difficulty (Balch, 1989; Bresnock, 1989; Dudycha & Carpenter, 1973; Frary, 1991; Hambleton & Traub, 1974; Kolstad, 1991; Oosterhof & Coats, 1984; Plake et al., 1982; Tinari, 1979; Tollefson, 1987). At UT, a multiple-choice test may consist of five types of items: usual multiple-choice, reason/assertation, case analysis, multiple selection or multiple multiple-choice, and diagram analysis. The changes in the composition of the number of each type of item in a UT test, and in the arrangement of the items including the order of the topics, cognitive domains, item format, and item difficulties may contribute to changes in item difficulty. In addition, the changes in the number of students and the achievement level of examinees also may alter the item difficulty (Allen & Yen, 1979; Brown, 1981; Tinari, 1979). As shown in Table 5, the number of students who took test 92.1 of this course (PING4448) is far fewer compared to the

number of students in the tests 87.1 and 89.1. The number of students in test 92.1 is only 91 while the number of students in test 87.1 and 89.1 are 239 and 123. Additional insight into the differences between FKIP old and new tests in the quality of item difficulty was available from data given in Table 5. The means of item difficulty of Physics and English courses in the new test 92.1 did not change very much from those in the old tests 87.1 and 89.1. Only two of six FKIP courses (33.3%), PING4441 (English for Education) and PING4448 (Business English), had a slightly higher mean of item difficulty in the new test 92.1 (The means were 0.553, 0.525, and 0.589 respectively for tests 87.1, 89.1, and 92.1 of PING4441; and were 0.619, 0.593, and 0.657 for tests 87.1, 89.1, and 92.1 of PING4448). However, as previously discussed, the course Business English (PING4448) did not show a higher percentage of acceptable items in the new test than was shown in the old tests. The course English for Education (PING4441) also did not show a higher percentage of acceptable items in the new test than was shown in the old tests.

Within FMIPA, the application of both Ebel's and UT's item difficulty criteria to the two old test administrations (88.1 and 89.1) in comparison to the new test administration (92.1), leads to the conclusion that

the distribution of item difficulties in its programs, Mathematics and Statistics, did not change very much (see Table 8). However, it appears from Table 8 that there was a higher percentage of acceptable items and moderate items, and a lower percentage of items judged to be difficult or very difficult ($p < 0.30$ under Ebel's criteria and $p \leq 0.40$ under UT's) in the new test administrations of the Statistics program. Meanwhile, in the Mathematics program, the percentage of acceptable items in the new test 92.1 judged to be acceptable under both Ebel's and UT's criteria was lower than those in the old tests 88.1 and 89.1. This result may be affected by the nature of items and students who took the program. These results could also be seen from the distribution of item difficulties in the courses. Although the distribution of item difficulties in the new test of Mathematics and Statistics courses did not change very much from those in the old tests, there is a shift in the number of acceptable, moderate, and hard items. A higher percentage of acceptable items and moderate items, and a lower percentage of items judged to be hard appeared in all Statistics courses. Meanwhile, in the two Mathematics courses used in this study, Calculus I (MATK4110) and Linear Algebra I (MATK4112), the percentages of items in the new test

judged to be acceptable under both Ebel's and UT's criteria were lower than those judged to be acceptable in the old tests 88.1 and 89.1. In Calculus I (MATK4110), the percentages of items judged to be acceptable under both Ebel's and UT's criteria were reduced respectively from 70.0% in test 88.1 to 66.7% in test 89.1, and to 56.7% in test 92.1; and from 100% in test 88.1 to 83.3% in test 89.1, and to 80.0% in test 92.1. In Linear Algebra I (MATK4112), the percentage of items judged to be acceptable under UT's criteria were also reduced from 87.5% in test 88.1 to 78.6% in test 89.1, and to 66.7% in test 92.1. Under Ebel's criteria, the percentages of acceptable items in this course were also lower in the new test 92.1 (41.7% in test 88.1, 53.6% in test 89.1, and 36.7% in test 92.1). These results may be affected by the nature of items and examinees. An item could be judged to be not acceptable under an evaluation criteria because it was too difficult or because the students have not yet learned the material (Brown, 1981). Meanwhile, the difficulty of an item is easily altered by rearranging test items or by testing students with different achievement levels (Allen & Yen, 1979; Hambleton & Traub, 1974; Tinari, 1979). Since the test 92.1 of Calculus I (MATK4110) consisted of 43.33% revised items and 56.67%

non-revised items, which were examined by an expert, a lower percentage of acceptable items in test 92.1 of this course might have occurred because of three factors: (1) the item revision did not appear to affect item difficulties; (2) the item arrangement changed; (3) the achievement level of examinees who took the old and new tests was different. In Linear Algebra I (MATK4112), on the other hand, since the test 92.1 of this course consisted of 93.33% non-revised items, a lower percentage of acceptable items in test 92.1 might have occurred because of changes in the item arrangement and in the achievement level of examinees. Further, it appears from Table 5 that the mean of item difficulty of Mathematics and Statistics courses in the new test 92.1 did not change a lot from those in the old tests 88.1 and 89.1 but all Statistics courses had a slightly higher mean of item difficulty. However, the new means were still low (less than 0.40). Meanwhile, none of the Mathematics courses had a higher mean of item difficulty in the new test. Of the two Mathematics courses, a great decrease in the mean of item difficulty occurred in the course, Linear Algebra I (MATK4112). In this course, the means were reduced from 0.378 in test 88.1 to 0.353 in test 89.1, and to 0.333 in test 92.1.

In summary, across the sample of three test

administrations, the Faculty of Education (FKIP) and the Faculty of Mathematics and Natural Science (FMIPA) have the same approximate quality in item difficulty. Most of their items met UT's criteria (about 80-90% items), but only about 50% of their items met Ebel's criteria. This means, that based on the comparisons with recommended difficulty offered by testing experts, about 50% of FKIP and FMIPA items failed to meet the item difficulty standard they offer. In addition, the important findings were that most of the items in FMIPA were hard (about 40% in Ebel's criteria and 70% in UT's) and most of FMIPA tests used in this study (86.7%) had low mean of item difficulty (less than 0.40). Most of the items in FKIP Physics program were also hard (about 30% in Ebel's criteria and 50% in UT's criteria) and 22.22% of the tests had low mean of item difficulty (less than 0.40).

Applying both Ebel's and UT's item difficulty criteria to the old test administrations in comparison to the new test administrations leads to the conclusion that the distribution of item difficulties in FKIP and FMIPA did not change very much. It seems that the revisions did not affect item difficulties. Within FKIP, the distribution of item difficulties in the new tests of Physics and English courses did not change very much so

as to have any impact on the measurement of student performance. Under Ebel's criteria, none of the Physics and English courses studied had a higher percentage of acceptable items in the new test than was found in the two old tests. In addition, only one course, English for Arts and Science (PING4447), had a lower percentage of items judged to be hard in the new test. Under UT's criteria, on the other hand, one of the three Physics courses, Introduction to Quantum Mechanics (PFIS4437), had a higher percentage of acceptable and moderate items, and had a lower percentage of items judged to be hard ($p \leq 0.40$). In this faculty, the great decrease occurred in the course, Business English (PING4448). In this course, the percentages of items judged to be acceptable under both Ebel's and UT's criteria were lower in the new test. In addition, the percentages of items judged to be moderate under UT's criteria were also lower in the new test. These results would be not positively affected by the evaluation criteria but by the nature of items and characteristics of the examinees. Within FMIPA, the distribution of item difficulties in the new tests of Mathematics and Statistics courses also did not change very much so as to have any impact on the measurement of student performance. A lower percentage of acceptable items

occurred in all Mathematics courses with the greatest decrease appearing in the course, Linear Algebra I (MATK4112). However, a higher percentage of acceptable items and moderate items, and a lower percentage of items judged to be hard ($p < 0.30$ under Ebel's criteria and $p \leq 0.40$ under UT's) occurred in all Statistics courses.

Comparison of revised or non-revised items in old and new tests. As shown in Table 9, the average item difficulty of the non-revised items in the old tests 87.1/89.1 in the Faculty of Education (FKIP) was 0.552, whereas in the new test 92.1, the average was 0.550. Thus, for non-revised items, the item difficulties remained stable overtime when the items were reused with different students (the change was only 0.002). It appears that the achievement of students in the two compared semester (87.1/89.1 and 92.1) was approximately the same. For revised items, the average item difficulty in the tests 87.1/89.1 was 0.494, whereas in the test 92.1, the average changed in the desired direction (an increase of 0.032 and still in the moderate category). This means that revisions in FKIP had the effect of modified item difficulty (increased the average item difficulty slightly [0.032]). The results of a paired t-test also shows that the average

Table 9
Average Item Difficulty within FKIP

	N of items	Old test	UT's criteria	New test	UT's criteria
FKIP					
Non-revised	178	0.552	Moderate	0.550	Moderate
Revised	160	0.494	Moderate	0.526	Moderate
Physics program					
Non-revised	72	0.459	Moderate	0.450	Moderate
Revised	76	0.386	Difficult	0.415	Moderate
PFIS4439					
Non-revised	23	0.479	Moderate	0.455	Moderate
Revised	26	0.526	Moderate	0.502	Moderate
PFIS4438					
Non-revised	22	0.463	Moderate	0.455	Moderate
Revised	28	0.341	Difficult	0.386	Difficult
PFIS4437					
Non-revised	27	0.438	Moderate	0.442	Moderate
Revised	22	0.280	Very Diff	0.349	Difficult
English program					
Non-revised	106	0.616	Moderate	0.617	Moderate
Revised	84	0.587	Moderate	0.626	Moderate
PING4447					
Non-revised	41	0.594	Moderate	0.597	Moderate
Revised	29	0.639	Moderate	0.647	Moderate

table continues...

	N of items	Old test	UT's criteria	New test	UT's criteria
PING4441					
Non-revised	30	0.652	Moderate	0.653	Moderate
Revised	30	0.469	Moderate	0.524	Moderate
PING4448					
Non-revised	35	0.610	Moderate	0.610	Moderate
Revised	25	0.667	Moderate	0.723	Moderate

item difficulties of FKIP revised items in the old tests 87.1/89.1 and new test 92.1 were significantly different, $t(159) = -2.40$, $p < .05$. The effectiveness of revisions in FKIP could also be seen from the average item difficulty in its programs and courses. In the Physics program, the average item difficulty of non-revised items tended to be stable overtime when the items were reused with different students (the average was only 0.009 lower in the new test and still in the moderate category). Meanwhile, the average item difficulty of revised items changed in the desired direction (the average changed from difficult in the old test 87.1/89.1 to moderate in the new test 92.1). However, the absolute value only changed minimally (only 0.029 higher). In the English program, the average item difficulty of non revised items also tended to be

stable overtime when the items were reused with different students (the average was only 0.001 higher in the new test 92.1). Meanwhile, the average item difficulty of revised items changed in the desired direction (the average was 0.039 higher in the new test 92.1 and still in the moderate category). These results reveal that achievement of students who took the old tests 87.1/89.1 and new tests 92.1 in FKIP Physics and English programs were approximately the same. Thus, revisions had the effect of modified item difficulty in FKIP Physics and English programs (increased the average item difficulty slightly [0.029 for Physics and 0.039 for English]). However, the results of paired t-tests show that, significant differences between the average item difficulties of revised items in the old tests 87.1/89.1 and new test 92.1 only occurred in English program ($t(83) = -2.01$, $p < .05$). Furthermore, Table 9 shows that the average item difficulty of revised items in two of the three Physics courses involved in this study (PFIS4438-Atomic Physics and PFIS4437-Introduction to Quantum Mechanics) and in all English courses (PING4447, English for Arts and Science; PING4441, English for Education; and PING4448, Business English) changed in the desired direction (the increase in the range of 0.008 to 0.069).

For non-revised items, the average item difficulty in these courses remained stable overtime when the items were reused with different students (the changes being in the range of 0.000 to 0.008). It appears that achievements of students who took the old tests 87.1/89.1 and new tests 92.1 of these courses were approximately the same. Thus, revisions in these courses had the effect of modified item difficulty (increased the average item difficulty slightly [the increase in the range of 0.008 to 0.069]).

However, shifts in the average item difficulty of Physics courses, PFIS4437 and PFIS4438, were not over the difficult category. In Introduction to Quantum Mechanics (PFIS4437), the average item difficulty of revised items shifted from very difficult to difficult while in Atomic Physics (PFIS4438), the average was still in the difficult category. These results might have appeared because writers or revisers believed that 'good' items equal hard items. Further improvement might be obtained by explaining that 'good' items do not equal hard items. Improvement might also be obtained by clarifying how writers or revisers are to change the items rather than give instruction to simply shift the item difficulty. For example, the writers may need to see a sample of items that might be defined as easy,

moderate, and hard.

Except for Business English (PING4448), the item revisions in Atomic Physics (PFIS4438); Introduction to Quantum Mechanics (PFIS4437); English for Arts and Science (PING4447); and English for Education (PING4441), also contributed to higher the percentage of items judged to be acceptable under UT's criteria (see Table 10). In those courses, the percentages of revised items judged to be acceptable under UT's criteria increased (the changes being in the range of 10% to 22.7%) while those of non-revised items did not. In Alternating Current (PFIS4439), on the other hand, the average item difficulty of both revised and non-revised items decreased by the same amount (0.024 lower in the new test 92.1). Thus, revisions in this course had no effect in increasing item difficulty. However, data given in Table 10 show that revisions in this course contributed to higher the percentage of acceptable items. The data show that the percentages of revised items in this course (PFIS4439) judged to be acceptable under either Ebel's or UT's criteria increased (the changes were 11.5% under Ebel's criteria and 3.9% under UT's), while those of non-revised items decreased (the changes were 8.7% under Ebel's criteria and 4.4% under UT's).

Table 10
Percentage of Acceptable Revised and Non-Revised Items within FKIP

	Percent of items in each category			
	Ebel's acceptable range: 0.30 - 0.70		UT's acceptable range: 0.20 - 0.90	
	Old test	New test	Old test	New test
Physics program				
PFIS4439				
Revised	57.7	69.2	84.6	88.5
Non-revised	56.5	47.8	95.7	91.3
PFIS4438				
Revised	53.6	46.4	67.9	78.6
Non-revised	81.8	72.7	90.9	86.4
PFIS4437				
Revised	36.4	36.4	68.2	90.9
Non-revised	81.5	74.1	88.9	88.9
English program				
PING4447				
Revised	24.1	37.9	58.6	72.4
Non-revised	51.2	46.3	97.6	95.1
PING4441				
Revised	43.3	56.7	76.7	86.7
Non-revised	50.0	56.7	83.3	83.3
PING4448				
Revised	40.0	32.0	80.0	60.0
Non-revised	51.4	48.6	91.4	88.6

Overall, in Table 10 it is shown that under Ebel's criteria, three of six courses in FKIP had a higher percentage of revised items judged to be acceptable, one course had no change, and two courses had a lower percentage. Meanwhile, for the non-revised items, only one course had a higher percentage of acceptable items and five courses had a lower percentage. Under UT's criteria, five courses had a higher percentage of revised items judged to be acceptable and only one course had a lower percentage. For the non-revised items, two courses had no change in the percentage of acceptable items and four courses had a lower percentage. Applying either Ebel's or UT's criteria, revisions in the majority of FKIP courses resulted in desired changes in the percentage of items which have acceptable difficulty.

In the Faculty of Mathematics and Natural Science (FMIPA), it appears that the average item difficulty of revised items changed in the desired direction (0.051 higher in the new tests 92.1) while the average item difficulty of non-revised items tended to be relatively more stable when the items were reused with different students (the average was only 0.012 higher in the new test 92.1). This means that item revisions in FMIPA had the effect of modified item difficulty (increased

Table 11
Average Item Difficulty within FMIPA

	N of items	Old test	UT's criteria	New test	UT's criteria
FMIPA					
Non-revised	115	0.372	Difficult	0.384	Difficult
Revised	34	0.321	Difficult	0.372	Difficult
Mathematics program					
Non-revised	44	0.370	Difficult	0.348	Difficult
Revised	15	0.374	Difficult	0.440	Moderate
MATK4110					
Non-revised	17	0.343	Difficult	0.350	Difficult
Revised	13	0.398	Difficult	0.475	Moderate
MATK4112					
Non-revised	27	0.387	Difficult	0.347	Difficult
Revised	2	0.224	Very Diff	0.215	Very Diff
Statistics program					
Non-revised	71	0.374	Difficult	0.407	Moderate
Revised	19	0.279	Very Diff	0.318	Difficult
STAT4110					
Non-revised	22	0.390	Difficult	0.415	Moderate
Revised	8	0.291	Very Diff	0.299	Very Diff
STAT4213					
Non-revised	24	0.381	Difficult	0.418	Moderate
Revised	6	0.254	Very Diff	0.259	Very Diff
STAT4334					
Non-revised	25	0.352	Difficult	0.389	Difficult
Revised	5	0.288	Very Diff	0.421	Moderate

the average item difficulty slightly [0.051]). The shift however, did not change the category of item difficulty level. After revisions, the average item difficulty in FMIPA was still categorized as difficult.

In FMIPA Mathematics program, the average item difficulty of revised items changed from the difficult category to the moderate category. However, the absolute value changed minimally (only 0.066 higher). For non-revised items, the average item difficulty decreased (0.022 lower) and remained stable in the difficult category when those items were reused with different students. These results reveal that revisions the in Mathematics program had the effect of modified item difficulty (increased the average item difficulty from difficult to moderate category).

Within this program, only the average item difficulty in Calculus I (MATK4110) changed in the desired direction. In this course, the average item difficulty of revised items changed from 0.398 (in the difficult category) to 0.475 (in the moderate category). Meanwhile, for non-revised items, the average item difficulty remained stable when the items were reused with different students (the average was only 0.007 higher in the new test 92.1 and still in the difficult category). Thus, the item revisions had the effect of

modified item difficulty. Item revisions in this course, however, did not appear to increase the percentage of acceptable items (see Table 12). In this course, the percentage of both revised and non-revised items judged to be acceptable under either Ebel's or UT's criteria did not increase (the percentages were the same or lower).

In Linear Algebra I (MATK4112), on the other hand, the average item difficulty of revised items remained relatively stable (the average was only 0.009 lower in the new test 92.1 and still in the very difficult category). Meanwhile, the average item difficulty of non-revised items decreased when those items were reused with different students (the average was 0.040 lower in the new test 92.1). This result indicates that achievement of students writing the new test 92.1 was slightly lower than achievement of students who wrote the old tests 88.1/89.1. Since the average item difficulty of revised items remained relatively stable when they were offered to students with lower achievements, it could be concluded that the item revisions had a slight positive effect on item difficulty. However, Table 12 indicates that revisions in this course did not appear to increase the percentage of acceptable items. In this course, the percentage of

Table 12
Percentage of Acceptable Revised and Non-Revised Items within FMIPA

	Percent of items in each category			
	Ebel's acceptable range: 0.30 - 0.70		UT's acceptable range: 0.20 - 0.90	
	Old test	New test	Old test	New test
Mathematics program				
MATK4110				
Revised	76.9	61.5	92.3	92.3
Non-revised	52.9	52.9	82.4	70.6
MATK4112				
Revised	0.0	0.0	50.0	50.0
Non-revised	48.1	40.7	88.9	70.4
Statistics program				
STAT4110				
Revised	25.0	37.5	100	87.5
Non-revised	68.2	63.6	95.5	95.5
STAT4213				
Revised	33.3	33.3	66.7	83.3
Non-revised	70.8	75.0	87.5	83.3
STAT4334				
Revised	40.0	60.0	60.0	60.0
Non-revised	36.0	56.0	76.0	84.0

revised items judged to be acceptable under either Ebel's or UT's criteria did not change, while that of non-revised items decreased (see Table 12). However, these results are limited since the number of revised items is small and far fewer than that of non-revised items. The number of revised items was only 2 (6.67%) while the number of non-revised items was 27 (93.33%). It is recognized that the average item difficulty is easily influenced by the number of items.

In Statistics program, it appears that the average item difficulty of revised items changed in the desired direction: increased from 0.279 (in the very difficult category) to 0.318 (in the difficult category). However, the average item difficulty of non-revised items also increased from 0.374 (in the difficult category) to 0.407 (in the moderate category) when those items were reused with different students.
Even though the average item difficulty of both revised and non-revised items moved from the difficult category to the moderate category, the absolute value did not change to any degree (only 0.039 higher for revised items and 0.033 for non-revised items). This means that revisions in this program had resulted in no meaningful change to item difficulty.

Within this program, only the average item

difficulty in Survey Sampling Method (STAT4334) changed in the desired direction. In this course, the average item difficulty of revised items changed from 0.288 (in the very difficult category) to 0.421 (in the moderate category). Meanwhile, the average item difficulty of non-revised items remained stable in the difficult category when those items were reused with different students (see Table 11). This means item revisions in this course resulted in desired changes in the average item difficulty. The revisions, however, did not appear to increase the percentage of acceptable items (see Table 12). In this course, the percentage of revised items judged to be acceptable under UT's criteria did not change, while that of non-revised items slightly increased (the shift was 8%). Under Ebel's criteria, the percentage of both revised and non-revised items judged to be acceptable increased by the same amount.

In Statistical Method I (STAT4110) and Applied Experimental Design (STAT4213), on the other hand, the average item difficulties of revised items tended to be stable overtime when those items were offered with different students. In Statistical Method I (STAT4110), the average item difficulty of revised items remained stable (0.008 higher in the new test 92.1 and still in the very difficult category). Meanwhile, the average

item difficulty of non-revised items changed from 0.390 (in the difficult category) to 0.415 (in the moderate category) when the items were reused with different students (see Table 12). Thus, item revisions in this course did not provide the desired changes in the average item difficulty. The revisions also did not appear to increase the percentage of items judged to be acceptable under UT's criteria. In this course, the percentage of revised items judged to be acceptable under UT's criteria was 12.5% lower in the new test while for the non-revised items, the percentage did not change. In Applied Experimental Design (STAT4213), the average item difficulty of revised items also tended to be stable (0.005 higher in the new test 92.1 and still in the very difficult category). Meanwhile, for non-revised items, the average item difficulty changed from 0.381 (in the difficult category) to 0.418 (in the moderate category). Revisions in this course then did not provide desired changes in the average item difficulty. The revisions also did not appear to increase the percentage of items judged to be acceptable under Ebel's criteria (see Table 12). In this course, the percentage of revised items judged to be acceptable under Ebel's criteria did not change while that of non-revised items increased (4.2% higher in the

new test 92.1). These results, however, are limited since the number of revised items in Statistical Method I (STAT4110) was only 8 (26.67%) while the number of non-revised items was 22 (73.37%). In Applied Experimental Design (STAT4213), the number of revised items was only 6 (20%) while the number of non-revised items was 24 (80%). It is recognized that the average item difficulty is affected by the number of items.

The results in the two courses, Statistical Method I (STAT4110) and Applied Experimental Design (STAT4213), might have occurred because (1) the writers or revisers did not follow the instructions given by coordinator of item revisions, (2) the writers or revisers believed that 'good' items equal hard items. Further improvement might be obtained by clarifying that 'good' items do not equal hard items. The improvement might also be obtained by clarifying how are writers or revisers change the items rather than giving instruction to simply shift the item difficulty. For example, the writers or revisers may need to see a sample of items that might be defined as easy, moderate, and hard.

Overall, in Table 12 it is shown that under Ebel's criteria, two of five courses in FMIPA had a higher percentage of revised items judged to be acceptable, two courses had no change, and one course had a lower

percentage. For non-revised items, two courses also had a higher percentage of acceptable items, one course had no change, and two courses had a lower percentage. Under UT's criteria, on the other hand, only one course had a higher percentage of revised items judged to be acceptable, three courses had no change, and one course had a lower percentage. For non-revised items, one course also had a higher percentage of acceptable items, one course had no change, and three courses had a lower percentage. Thus, using either Ebel's or UT's criteria, item revisions in FMIPA courses had resulted in no meaningful change to the percentage of acceptable items.

In summary, item revisions in both faculties, the Faculty of Education (FKIP) and the Faculty of Mathematics and Natural Science (FMIPA) resulted in desired changes in the item difficulty of most of their tests.

In FKIP, it seems that the item revisions in two of three Physics courses (PFIS4438-Atomic Physics and PFIS4437-Introduction to Quantum Mechanics) and in two of three English courses (PING4447-English for Arts and Science and PING4441-English for Education) resulted in desired changes in both the average item difficulty and the percentage of acceptable items. In those courses, the average item difficulties of revised items increased

with the changes in a range of 0.008 to 0.069 while those of non-revised items remained stable (the changes in a range of 0.000 to 0.008). In those courses, the percentages of revised items judged to be acceptable under UT's criteria also increased with the changes in a range of 10% to 22.7% while those of non-revised items did not.

In Alternating Current (PFIS4439), item revisions had no effect in increasing item difficulty. However, the item revisions contributed to higher the percentage of acceptable items. In this course, the percentage of revised items increased (the changes were 11.5% under Ebel's criteria and 3.9% under UT's), while that of non-revised items decreased (the changes were 8.7% under Ebel's criteria and 4.4% under UT's).

In Business English (PING4448), on the other hand, the item revisions did not increase the percentage of acceptable items but they contributed to higher the average item difficulty. In this course, the average item difficulty of revised items was 0.056 higher in the new test 92.1, while that of non-revised items did not change.

In FMIPA, of the two Mathematics courses, only item revisions in Calculus I (MATK4110) resulted in desired changes in the average item difficulty. In this course,

the average item difficulty of revised items shifted from difficult to moderate, while that of non-revised items tended to be stable (still in the difficult category). However, the item revisions in this course did not appear to increase the percentage of acceptable items.

In the course, Linear Algebra I (MATK4112), it seems that the item revisions provide a slight positive effect on item difficulty. However, the revisions did not appear to increase the percentage of acceptable items. Of the three Statistics courses, only item revisions in Survey Sampling Method (STAT4334), resulted in desired changes in the average item difficulty. In this course, the average item difficulty of revised items changed from very difficult to moderate, while that of non-revised items remained stable in the difficult category. However, the item revisions in this course did not appear to increase the percentage of acceptable items. In this course, the percentages of both revised and non-revised items judged to be accepted under Ebel's criteria increased by the same amount. Under UT's criteria, the percentage of revised items judged to be acceptable did not change while that of non-revised items slightly increased (the shift was 8%).

In Statistical Method I (STAT4110) and Applied

Experimental Design (STAT4213), the revisions did not provide desired changes in either the average item difficulty or the percentage of acceptable items. In these courses, the average item difficulty of revised items was still in the very difficult category, while that of non-revised items shifted from difficult to moderate.

In addition, the percentage of revised items in STAT4110 judged to be acceptable under UT's criteria was 12.5% lower in the new test 92.1 while that of non-revised items did not change. In STAT4213, the percentage of revised items judged to be acceptable under Ebel's criteria did not change while that of non-revised items was 4.2% lower in the new test 92.1.

Item Discrimination

In order to judge whether item discrimination in both faculties, the Faculty of Education (FKIP) and the Faculty of Mathematics and Natural Science (FMIPA), had changed and to determine whether the quality of item discrimination between faculties was different, two steps were taken. First, Ebel's criteria for item discrimination were applied to the old (87.1, 88.1, and 89.1) and new (92.1) tests to determine if the distribution of item discrimination had changed. Differences in the mean of item discrimination and in

the percentage of items which meet the criteria for item discrimination were analyzed, and faculty-by-faculty, program-by-program, and course-by-course comparisons were conducted. Second, Ebel's criteria for item discrimination were applied to the revised and non-revised items on both old and new tests to determine the effects of item revision upon item discrimination. The effects of item revision upon item discrimination were analyzed through the analysis of the differences in the individual item discrimination, the average item discrimination, and the percentage of items judged to be good and poor. Faculty-by-faculty, program-by-program, and course-by-course comparisons were conducted.

Comparison of old and new tests. As shown in Table 13, across the sample of 3 test administrations, the Faculty of Education (FKIP) was slightly better in the quality of item discrimination than the Faculty of Mathematics and Natural Science (FMIPA). The percentage of poor items (items with discrimination less than 0.20) in FKIP was the same as that in FMIPA (about 30-40% items) but the percentage of good/very good items (items with discrimination equal to or higher than 0.30) in FKIP was slightly higher than that in FMIPA (about 40-50% items for FKIP and 30-40% items for FMIPA). Table 13 also shows that, across the three test

Table 13
Distribution of Individual Item Discrimination between Faculties

Discrimination range	Percent of items in each category					
	FKIP			FMIPA		
	87.1	89.1	92.1	88.1	89.1	92.1
< 0.00	7.7	4.7	5.9	0.7	3.4	8.7
0.00 - 0.099	9.8	8.7	6.5	7.2	11.7	11.3
0.10 - 0.199	17.0	21.6	20.4	21.7	22.8	18.0
0.20 - 0.299	22.6	22.6	26.0	34.8	24.8	24.7
0.30 - 0.399	24.1	22.6	26.3	21.0	21.4	16.0
> 0.399	18.8	19.7	14.8	14.5	15.9	21.3

administrations, 60 to 70% items in both FKIP and FMIPA met UT's acceptable values in item discrimination (Discrimination > 0.20). However, considering Ebel's criteria concerning a good test, none of the tests in either faculty used in this study met the proportion of item discrimination recommended by Ebel (see Table 5). In terms of discrimination, Ebel (1972) suggested that a good test should have an appropriate proportion in discrimination indices (Over 25% of items should have values of 0.40 and up, less than 25% of items should have values in a range of 0.20 to 0.39, less than 15% should have values in a range of 0.01 to 0.19, and less than 5% have zero or negative values). In fact, most of

UT's tests consist of over 25% of items having discrimination values in a range of 0.20 to 0.39 (marginal and reasonably good items) and over 15% having discrimination values in a range of 0.01 to 0.19 (poor items) with the average discrimination classified as marginal. Overall, Table 13 represented that, in both faculties, the distribution of item discrimination in the new test administration (92.1) did not change very much from those in the two old test administrations (87.1 and 89.1 for FKIP and 88.1 and 89.1 for FMIPA).

In the Faculty of Education (FKIP), the changes were only 1.7% and 2.2% for the percentage of poor items (items with discrimination less than 0.20) and, 1.8% and 1.2% for the percentage of good/very good items (items with discrimination equal to or higher than 0.30). The percentage of items judged to be marginal (discrimination in a range of 0.20 to 0.299) was 3.4 % higher than the two old tests 87.1 and 89.1. In this faculty, although the percentage of poor items (items with discrimination less than 0.20) was slightly lower in the new test 92.1, the percentage of good/very good items (items with discrimination equal to or higher than 0.30) was also slightly lower.

In the Faculty of Mathematics and Natural Science (FMIPA), the changes were also only 8.4% and 0.1% for

the percentage of poor items, 1.8% and 0% for the good/very good items, and 10.1% and 0.1% for the marginal items. In this faculty, two points are noteworthy. First, the percentage of poor items in the new test 92.1 was 8.4% higher than in the test of 88.1 and 0.1% higher than in the test of 89.1. Second, the percentage of good/very good items was 1.8% higher in the new test 92.1 than in the test of 88.1 and was the same as that in the test of 89.1.

Additional insight into the differences between item discrimination of old and new tests in the Faculty of Mathematics and Natural Science (FMIPA) and the Faculty of Education (FKIP) are available from the data given in Tables 14. Within FKIP, the distribution of item discrimination in the new test administration of Physics and English programs did not change very much from those in the two old test administrations (see Table 14). In the Physics program, the changes were 7% and 10.5% for the percentage of poor items (discrimination less than 0.20), 8.9% and 7.2% for the marginal items (discrimination in a range of 0.20 to 0.299), and 2% and 2.7% for good/very good items (discrimination equal to or higher than 0.30). In the English program, the changes were 1.1% and 3.7% for poor items, 0.5% and 0.5% for marginal items, and 0.6% and

Table 14
Distribution of Individual Item Discrimination within FKIP

Discrimination range	Physics program	English program	PFIS 4439	PFIS 4438	PFIS 4437	PING 4447	PING 4441	PING 4448
	Percent of items in each category							
87.1								
< 0.00	8.5	7.2	5.0	9.1	11.1	5.7	9.0	7.1
0.00 - 0.099	11.6	8.7	12.5	11.4	11.1	5.7	11.9	8.6
0.10 - 0.199	23.3	13.0	20.0	31.8	17.8	11.4	16.4	11.4
0.20 - 0.299	20.2	24.2	17.5	25.0	17.8	24.3	25.4	22.9
0.30 - 0.399	24.8	23.7	37.5	13.6	24.4	25.7	17.9	27.1
> 0.399	11.6	23.2	7.5	9.1	17.8	27.1	19.4	22.9
89.1								
< 0.00	5.0	4.5	0.0	8.3	6.0	5.0	1.3	8.3
0.00 - 0.099	11.3	6.8	12.0	10.0	12.0	11.3	5.0	3.3
0.10 - 0.199	30.6	15.0	26.0	30.0	36.0	21.3	15.0	6.7
0.20 - 0.299	21.9	23.2	26.0	15.0	26.0	27.5	23.8	16.7
0.30 - 0.399	21.9	23.2	24.0	23.3	18.0	21.3	22.5	26.7
> 0.399	9.4	27.3	12.0	13.3	2.0	13.8	32.5	38.3
92.1								
< 0.00	7.4	4.7	10.2	8.0	4.1	5.7	1.7	6.7
0.00 - 0.099	8.1	5.3	2.0	14.0	8.2	5.7	6.7	3.3
0.10 - 0.199	20.9	20.0	14.3	30.0	18.4	24.3	13.3	21.7
0.20 - 0.299	29.1	23.7	26.5	32.0	28.6	20.0	25.0	26.7
0.30 - 0.399	27.0	25.8	34.7	14.0	32.7	32.9	21.7	21.7
> 0.399	7.4	20.5	12.2	2.0	8.2	11.4	31.7	20.0

4.2% for good/very good items. A slight improvement in item discrimination occurred in one program, the Physics program. In this program, the percentage of items judged as poor was lower in the new test 92.1, while in the other program the percentage was not. However, the percentages of items judged as good/very good in both programs were not higher in the new test 92.1.

Of three courses in the Physics program, improvement in item discrimination took place in two courses, Alternating Current (PFIS4439) and Introduction to Quantum Mechanics (PFIS4437), with the highest improvement occurring in Alternating Current. In Alternating Current (PFIS4439), the percentage of poor items was lower in the new test 92.1 (11% lower than the percentage in test 87.1 and 11.5% lower than the percentage in test 89.1), while the percentage of good/very good items was slightly higher (1.9% higher than the percentage in test 87.1 and 10.9% higher than the percentage in test 89.1). The percentage of very good items in the new test 92.1 was also slightly higher than the percentage in old tests 87.1 and 89.1 (respectively, 4.7% and 0.2% higher). In Introduction to Quantum Mechanics (PFIS4437), the percentage of poor items was also lower in the new test 92.1 (9.3% lower than the percentage in test 87.1 and 13.3% lower than

the percentage in test 89.1), while the percentage of good items in the new test 92.1 was higher than the percentage of test 89.1 (20.9% higher) but 1.3% lower than the percentage of test 87.1. In Atomic Physics (PFIS4438), on the other hand, the percentage of poor items in the new test 92.1 was only 0.3% lower than the percentage of test 87.1 and 3.7% higher than the percentage of test 89.1 while the percentage of good/very good items was lower in the new test 92.1 (6.7% lower than the percentage in test 87.1 and 20.6% lower than the percentage in test 89.1). In addition, the percentage of very good items (items with discrimination higher than 0.399) in this course decreased (the percentages were 9.1% in test 87.1, 13.3% in test 89.1, and 2.0% in test 92.1). Data provided in Table 5 also supported these results. Only the courses, Alternating Current and Introduction to Quantum Mechanics, had a higher mean of item discrimination in the new test. However, the new means were still categorized as marginal. These results could be affected by the nature of items and students. At UT, a multiple-choice test may consist of five types of items: usual multiple-choice, reason/assertation, case analysis, multiple selection or multiple multiple-choice, and diagram analysis. The changes in the number

of items and in the composition of the number of each type of items may contribute to the changes in item discrimination. As shown in Table 5, the percentages of items retained in test 92.1 of the courses, Alternating Current (PFIS4439) and Introduction to Quantum Mechanics (PFIS4437) were 98, while that of Atomic Physics (PFIS4438) was 100. The elimination of some items from the tests may affect the item discrimination (Cureton, 1950). In addition, the differences in the number of students who took the tests may contribute to the changes (Allen & Yen, 1979; Ebel & Frisbie, 1986).

Within the English program, none of the three courses had a lower percentage of poor items (discrimination less than 0.20) in the new test 92.1. In addition, none of the courses had a higher percentage of good/very good items (discrimination equal to or higher than 0.30). However, in comparison to the Physics courses, the percentages of good/very good items and very good items (discrimination higher than 0.399) of the English courses in every test administration used in this study were higher than those of the Physics courses (see Table 14). These results are surprising. Because of the nature of science courses which provide an obvious correct or incorrect answer for test items, it is recognized that it is easier to construct

discriminating items for science courses than those for non-science (social) courses. However, mastery of the subject matter and skill in test construction on the part of the item writers play a large role in constructing unambiguous stems and options; and consequently, play a large role in developing discriminating items. An ambiguous item could occur as a result of an ambiguous idea in the mind of the item writer.

Within the Faculty of Mathematics and Natural Science (FMIPA), the distribution of item discriminations in the new test administration of the Mathematics and Statistics programs also did not change very much from those in the two old test administrations 88.1 and 89.1. However, as in FKIP, a slight improvement in item discrimination took place in one program, the Statistics program. In this program, the percentage of items judged as poor (items with discrimination less than 0.20) was lower in the new test 92.1, while the percentage of very good items (items with discrimination higher than 0.399) was higher (see Table 15). Meanwhile, in the Mathematics program, the percentage of poor items was higher in the new test 92.1 (respectively, 23.6% and 9.1% higher than the percentage of test 88.1 and 89.1). In this program, the

Table 15
Distribution of Individual Item Discrimination within FMIPA

Discrimination range	Math program	Stat program	MATK 4110	MATK 4112	STAT 4110	STAT 4213	STAT 4334
		Percent of items in each category					
88.1							
< 0.00	0.0	1.2	0.0	0.0	0.0	3.3	0.0
0.00 - 0.099	7.4	7.1	6.7	8.3	3.4	10.0	8.0
0.10 - 0.199	7.4	31.0	6.7	8.3	24.1	23.3	48.0
0.20 - 0.299	22.2	42.9	26.7	16.7	48.3	40.0	40.0
0.30 - 0.399	25.9	17.9	33.3	16.7	24.1	23.3	4.0
> 0.399	37.0	0.0	26.7	50.0	0.0	0.0	0.0
89.1							
< 0.00	1.7	4.6	3.3	0.0	3.3	10.0	0.0
0.00 - 0.099	6.9	14.9	13.3	0.0	13.3	20.0	11.1
0.10 - 0.199	20.7	24.1	20.0	21.4	33.3	20.0	18.5
0.20 - 0.299	27.6	23.0	30.0	25.0	26.7	16.7	25.9
0.30 - 0.399	27.6	17.2	23.3	32.1	13.3	20.0	18.5
> 0.399	15.5	16.1	10.0	21.4	10.0	13.3	25.9
92.1							
< 0.00	6.7	10.0	6.7	6.7	3.3	16.7	10.0
0.00 - 0.099	10.0	12.2	10.0	10.0	10.0	13.3	13.3
0.10 - 0.199	21.7	15.6	16.7	26.7	20.0	10.0	16.7
0.20 - 0.299	18.3	28.9	20.0	16.7	46.7	30.0	10.0
0.30 - 0.399	20.0	13.3	13.3	26.7	16.7	13.3	10.0
> 0.399	23.3	20.0	33.3	13.3	3.3	16.7	40.0

percentage of good/very good items (items with discrimination equal to or higher than 0.30) in the new test 92.1 was 19.6% lower than the percentage in test 88.1 and only 0.2% higher than the percentage in test 89.1. The results could also be seen from the distribution of item discrimination in FMIPA courses. It appears that three of five courses (one Mathematics course, Calculus I or MATK4110; and two Statistics courses, Applied Experimental Design-STAT4213 and Survey Sampling Method-STAT4334) had a higher percentage of very good items in the new test 92.1. However, none of the courses studied had a lower percentage of poor items in the new test than those in the two old tests. In addition, the percentages of items which have negative discrimination were increased in the new test of all courses involved in this study. In the test 92.1 of Linear Algebra I (MATK4112) which consists of 6.67% revised and 93.33% non-revised items drawn from both tests 88.1 and 89.1 and in test 92.1 of Survey Sampling Method (STAT4334) which consists of 17.67% revised and 82.33% non-revised items, the percentages of items that have negative discrimination increased respectively from 0% to 6.7% and 0% to 10% (see Table 15). In Linear Algebra I (MATK4112), the percentage of good/very good items (items with discrimination equal to or higher than

0.30) also decreased (respectively, 26.7% and 13.5% lower than the percentages of tests 88.1 and 89.1). Meanwhile, in Survey Sampling Method (STAT4334), the percentage of good/very good items increased (respectively, 46% and 5.6% higher than the percentages of tests 88.1 and 89.1). These results indicate that test revisers need to reconsider some of the revised and non-revised items in the new test 92.1 of these courses, especially for Linear Algebra I (MATK4112). The items probably were still ambiguous and have incorrect key options.

In summary, across the sample of three test administrations, FKIP was slightly better in the quality of item discrimination than FMIPA. The percentage of poor items in FKIP was the same as those in FMIPA but the percentage of good/very good items in FKIP was slightly higher than those in FMIPA. However, considering Ebel's criteria concerning a good test, none of the tests in either faculty met Ebel's criteria. Based on comparisons of old and new tests, overall, the distribution of item discrimination in both faculties did not change very much. The results were mixed. On the one hand, only two of six courses in FKIP (Alternating Current- PFIS4439 and Introduction to Quantum Mechanics-PFIS4437) had a lower percentage of

poor items in the new test 92.1 and none of the courses in FMIPA had a lower percentage. On the other hand, three of five courses in FMIPA (Calculus I-MATK4110, Applied Experimental Design-STAT4213, and Survey Sampling Method-STAT4334) had a slightly higher percentage of very good items and one of six courses in FKIP (Alternating Current-PFIS4439) had a higher percentage of very good items.

Comparison of revised or non-revised items in old and new tests. Appendices A, B, C, and D show the item discrimination indices on the items used without revision (non-revised items) on both the first and second use, and the item discrimination indices on revised items on both before and after revision in each program of the two faculties, Faculty of Education (FKIP) and Faculty of Mathematics and Natural Science (FMIPA).

It appears from Appendix A that item revisions in the two of three Physics courses in FKIP, Alternating Current (PFIS4439) and Introduction to Quantum Mechanics (PFIS4437), resulted in desired changes in the individual item discrimination.

In the course, Alternating Current (PFIS4439), 17 of 26 revised items (65%) showed an improvement in discrimination with the lowest increase being 0.023 and

the greatest being 0.323. Meanwhile, only 10 of 23 non-revised items (43%) improved in discrimination with the lowest increase being 0.037 and the greatest being 0.228. Only 9 of 26 revised items (35%) showed a decrease in discrimination with the lowest decrease being 0.019 and the greatest being 0.486. Meanwhile, 13 of 23 non-revised items (57%) decreased in discrimination with the lowest decrease being 0.017 and the highest decrease being 0.463.

In the course, Introduction to Quantum Mechanics (PFIS4437), 13 of 22 revised items (59%) showed an improvement in discrimination with the lowest increase being 0.032 and the highest increase being 0.410 (changed the value from negative to positive). Meanwhile, only 11 of 27 non-revised items (41%) improved in discrimination with the lowest increase being 0.009 and the greatest being 0.164. Only 9 of 22 revised items (41%) showed a decrease in discrimination with the lowest decrease being 0.057 and the greatest being 0.179. Meanwhile, 16 of 27 non-revised items (59%) decreased in discrimination with the lowest decrease being 0.036 and the highest decrease being 0.356.

These results were also supported by the data given in Tables 16 and 17. Table 16 shows that, in the two

Table 16
Average Item Discrimination within FKIP

	N of items	Old test	Ebel's criteria	New test	Ebel's criteria
FKIP					
Non-revised	178	0.327	R. Good	0.283	Marginal
Revised	160	0.198	Poor	0.229	Marginal
Physics program					
Non-revised	72	0.312	R. Good	0.264	Marginal
Revised	76	0.179	Poor	0.199	Poor
PFIS4439					
Non-revised	23	0.307	R. Good	0.274	Marginal
Revised	26	0.222	Marginal	0.249	Marginal
PFIS4438					
Non-revised	22	0.286	Marginal	0.246	Marginal
Revised	28	0.152	Poor	0.135	Poor
PFIS4437					
Non-revised	27	0.336	R. Good	0.269	Marginal
Revised	22	0.162	Poor	0.221	Marginal
English program					
Non-revised	106	0.337	R. Good	0.296	Marginal
Revised	84	0.216	Marginal	0.256	Marginal
PING4447					
Non-revised	41	0.349	R. Good	0.275	Marginal
Revised	29	0.174	Poor	0.225	Marginal

Note. R. Good is Reasonably Good

table continues...

	N of items	Old test	Ebel's criteria	New test	Ebel's criteria
PING4441					
Non-revised	30	0.329	R. Good	0.346	R. Good
Revised	30	0.175	Poor	0.281	Marginal
PING4448					
Non-revised	35	0.329	R. Good	0.279	Marginal
Revised	25	0.312	R. Good	0.262	Marginal

Note. R. Good is Reasonably Good

Table 17
Percentage of Good and Poor Items on Revised and Non-Revised Items within FKIP

	Percent of items in each category			
	Very good and Reasonably good items (Discr ≥ 0.30)		Poor items (Discr < 0.20)	
	Old Test	New Test	Old test	New Test
Physics program				
PFIS4439				
Revised	34.6	46.2	50.0	34.6
Non-revised	65.2	47.8	17.4	17.4
PFIS4438				
Revised	7.1	10.7	78.6	71.4
Non-revised	54.5	22.7	13.6	27.3
PFIS4437				
Revised	18.2	31.8	63.6	36.4
Non-revised	63.0	48.1	11.1	25.9

table continues ...

	Percent of items in each category			
	Very good and Reasonably good items (Discr ≥ 0.30)		Poor items (Discr < 0.20)	
	Old test	New test	Old test	New test
English program				
PING4447				
Revised	10.3	34.5	58.6	44.8
Non-revised	68.3	51.2	7.3	29.3
PING4441				
Revised	16.7	40.0	60.0	33.3
Non-revised	53.3	66.7	10.0	10.0
PING4448				
Revised	56.0	36.0	20.0	36.0
Non-revised	65.7	45.7	14.3	28.6

courses, the average discrimination of revised items changed from 0.222 to 0.249 (still in the marginal category) for the course of Alternating Current (PFIS4439) and shifted from 0.162 (poor) to 0.221 (marginal) for the course Introduction to Quantum Mechanics (PFIS4437). Meanwhile, for the non-revised items, the average item discrimination in the two courses decreased (0.033 lower for Alternating Current-PFIS4439 and 0.067 lower for Introduction to Quantum Mechanics-PFIS4437). In Table 17, it is indicated that in the two courses, the percentage of revised items

judged to be good/very good increased (11.6% higher for Alternating Current-PFIS4439 and 13.6% higher for Introduction to Quantum Mechanics-PFIS4437) while the percentage of revised items judged to be poor decreased (15.4% lower for Alternating Current-PFIS4439 and 27.2% lower for Introduction to Quantum Mechanics-PFIS4437). Meanwhile, for the non-revised items, the percentages of good/very good items in the two courses decreased (17.4% lower for Alternating Current-PFIS4439 and 14.9% lower for Introduction to Quantum Mechanics-PFIS4437) and the percentages of poor items were not reduced (did not change for Alternating Current-PFIS4439 and 14.8% higher for Introduction to Quantum Mechanics-PFIS4437).

In Atomic Physics (PFIS4438), on the other hand, the average discrimination of both revised and non-revised items decreased (0.017 lower for revised items and 0.040 lower for non-revised items). The item revisions then did not provide desired changes in the average item discrimination. However, Table 17 indicates that the item revisions resulted in desired changes in the percentage of good/very good and poor items. In this course, the percentage of revised items judged to be good/very good increased (3.6% higher) while the percentage of items judged to be poor decreased (7.2% lower). Meanwhile, for the non-revised items, the

the percentage of items judged to be good/very good decreased (31.8% lower) and the percentage of items judged to be poor increased (13.7% higher).

Furthermore, in Appendix B it is shown that item revisions in two of three English courses in the Faculty of Education (FKIP), English for Arts and Science (PING4447) and English for Education (PING4441), resulted in desired changes in the individual item discrimination.

In the course, English for Arts and Science (PING4447), 18 of 29 revised items (62%) showed improvement in discrimination with the lowest increase being 0.007 and the highest being 0.551 (the value changed from negative to positive). Meanwhile, only 12 of 41 non-revised items (29%) improved in discrimination with the lowest increase being 0.003 and the highest being 0.244. Only 11 of 29 revised items (38%) showed a decrease in discrimination with the lowest decrease being 0.010 and the greatest being 0.232. Meanwhile, 29 of 41 non-revised items (71%) decreased in discrimination with the lowest decrease being 0.015 and the highest being 0.383.

In English for Education (PING4441), 21 of 30 revised items (70%) showed improvement in discrimination with the lowest increase being 0.005 and

the highest being 0.624 (the value changed from negative to positive). Meanwhile, only 11 of 30 non-revised items (37%) improved in discrimination with the lowest increase being 0.008 and the highest being 0.209. Only 9 of 30 revised items (30%) showed a decrease in discrimination with the lowest decrease being 0.002 and the highest being 0.336. Meanwhile, there were 18 of 30 non-revised items (60%) decreased in discrimination with the lowest decrease being 0.008 and the highest being 0.176. One non-revised item had no change in discrimination.

The effect of item revisions on item discrimination in the two courses could also be seen from data given in Tables 16 and 17. In the two courses (see Table 16), the average discrimination of revised items increased from the poor to marginal category (the average was 0.051 higher for PING4447-English for Arts and Science and 0.106 higher for PING4441-English for Education). However, the results of paired t-tests show that, significant differences between the average discrimination of revised items in the old and new tests only occurred in English for Education (PING4441) ($\underline{t}(29) = -2.94$, $\underline{p} < .01$).

Meanwhile, for the non-revised items, the average discrimination in English for Arts and Science

(PING4447) decreased (0.074 lower) and in English for Education (PING4441) slightly increased (0.017 higher). Table 17 also shows that, in the two courses, the percentages of revised items judged to be good/very good increased (24.2% higher for PING4447-English for Arts and Science and 23.3% higher for PING4441-English for Education) while the percentages of revised items judged to be poor decreased (13.8% lower for PING4447-English for Arts and Science and 26.7% lower for PING4441-English for Education). Meanwhile, for the non-revised items, the percentage of good/very good items in English for Arts and Science (PING4447) was 17.1% lower in the new test while the percentage in English for Education (PING4441) was 13.4% higher. In the two courses, the percentages of non-revised items judged to be poor were not reduced (22% higher for PING4447-English for Arts and Science and did not change for PING4441-English for Education).

In Business English (PING4448), on the other hand, the average discrimination of both revised and non-revised items decreased (0.050 lower for both revised and non-revised items). In addition, the percentage of both revised and non-revised items judged to be good/very good decreased (20% lower for both revised and non-revised items). Meanwhile, the percentage of both

revised and non-revised items judged to be poor increased (16% higher for revised items and 14.3% higher for non-revised items). These results indicate that item revisions in this course did not provide desired changes in the item discrimination. These results might have occurred because (1) the writers or revisers did not follow the instruction given by the coordinators of item revisions, (2) the writers or revisers did not know how to produce a more discriminating item. Further improvement might be obtained by clarifying how are the writers or revisers change the items so that the items may have a higher discrimination. For example, the writers or revisers may need to recognize the items that might be defined as good (have high discrimination) and poor (have low discrimination). Overall, Table 16 shows that, item revisions in the Faculty of Education (FKIP) resulted in desired changes in the average discrimination. In this faculty, the average discrimination of revised items was 0.031 higher in the new test while the average of non-revised items was 0.044 lower. The result of a paired t-test also indicates that the average discrimination of revised items in the old and new tests were significantly different, $\underline{t}(159) = -2.35$, $\underline{p} < .05$. In its programs, item revisions also resulted in desired changes in the

item discrimination. In these programs, the average discrimination of revised items increased (respectively, 0.020 and 0.040 higher for Physics and English) while the averages of non-revised items decreased (respectively, 0.048 and 0.041 lower for Physics and English). However, the results of paired t-tests show that, significant differences between the average discrimination of revised items in the old and new tests only appeared in English program ($\underline{t}(83) = -2.14$, $\underline{p} < .05$).

As shown in Appendix C, item revisions in the Faculty of Mathematics and Natural Science (FMIPA) did not provide a lot of contribution to changes in individual item discrimination. In the course of Calculus I (MATK4110), 8 of 13 revised items (62%) showed improvement in discrimination with the lowest increase being 0.033 and the greatest being 0.358. Meanwhile, 14 of 17 non-revised items (82%) improved in discrimination with the lowest increase being 0.009 and the highest being 0.283. In the course, Linear Algebra I (MATK4112), 1 of 2 revised items showed improvement in discrimination, while 10 of 27 non-revised items (37%) improved in discrimination. However, Table 18 shows that item revisions in Mathematics program resulted in desired changes in the average item discrimination. As

shown in Table 18, the average discrimination of revised items in this program was 0.038 higher in the new test while the average of non-revised items was 0.034 lower. Within this program, however, only item revisions in Calculus I (MATK4110) resulted in desired changes in the average discrimination and the percentages of items judged to be good/very good and poor (see Tables 18 and 19). As shown in Table 18, the average discrimination of revised items in this course was 0.056 higher in the new test while the average of non-revised items was only 0.038 higher. However, the average item discrimination of revised items was still in the marginal category. Table 19 shows that for the revised items in this course, the percentage of good/very good items was 15.4% higher in the new test while the percentage of poor items was 23% lower. For non-revised items, the percentage of good/very good items was only 5.8% higher in the new test while the percentage of poor items was only 5.9% lower.

In Linear Algebra I (MATK4112), on the other hand, item revisions did not provide desired changes in either the average item discrimination or the percentage of good/very good and poor items. In this course, the average discrimination of both revised and non-revised items decreased (0.077 lower for revised items and 0.078

Table 18
Average Item Discrimination within FMIPA

	N of items	Old test	Ebel's criteria	New test	Ebel's criteria
FMIPA					
Non-revised	115	0.277	Marginal	0.267	Marginal
Revised	34	0.197	Poor	0.206	Marginal
Mathematics program					
Non-revised	44	0.308	R. Good	0.274	Marginal
Revised	15	0.206	Marginal	0.244	Marginal
MATK4110					
Non-revised	17	0.276	Marginal	0.314	R. Good
Revised	13	0.210	Marginal	0.267	Marginal
MATK4112					
Non-revised	27	0.327	R. Good	0.249	Marginal
Revised	2	0.174	Poor	0.097	Poor
Statistics program					
Non-revised	71	0.258	Marginal	0.262	Marginal
Revised	19	0.190	Poor	0.176	Poor
STAT4110					
Non-revised	22	0.257	Marginal	0.234	Marginal
Revised	8	0.218	Marginal	0.174	Poor
STAT4213					
Non-revised	24	0.240	Marginal	0.242	Marginal
Revised	6	0.154	Poor	0.152	Poor
STAT4334					
Non-revised	25	0.227	Marginal	0.306	R. Good
Revised	5	0.189	Poor	0.208	Marginal

Note. R. Good is Reasonably Good

Table 19
Percentage of Good and Poor Items on Revised and Non-Revised Items within FMIPA

	Percent of items in each category			
	Very good and Reasonably good items (Discr ≥ 0.30)		Poor items (Discr < 0.20)	
	Old test	New test	Old test	New test
Mathematics program				
MATK4110				
Revised	23.1	38.5	61.5	38.5
Non-revised	47.1	52.9	35.3	29.4
MATK4112				
* Revised	50.0	0.0	50.0	100
Non-revised	55.6	44.4	25.9	37.0
Statistics program				
STAT4110				
Revised	12.5	0.0	50.0	62.5
Non-revised	36.4	27.3	22.7	22.7
STAT4213				
Revised	0.0	33.3	66.7	66.7
Non-revised	37.5	29.2	29.2	33.3
STAT4334				
Revised	0.0	40.0	60.0	60.0
Non-revised	32.0	52.0	28.0	36.0

Note. *The number of revised items was only 2 while the number of non-revised items was 27

lower for non-revised). In addition, the percentages of both revised and non-revised items judged to be good/very good decreased. Before revision, 1 of 2 revised items (50%) was categorized as good item but after revision (on the new test), there was no good item (see Appendix C). For the non-revised items, 15 of 27 items (55.6%) in the old tests (first use) were categorized as good/very good while in the new test (second use), only 12 of 27 items (44.4%) were categorized as good/very good. Meanwhile, the percentages of both revised and non-revised items judged to be poor increased. Before revision (in the old test), 1 of 2 revised items (50%) was classified as poor item but after revision (in the new test), there were 2 poor items (100%) (see Appendix C). For the non-revised items, only 7 of 27 items (25.9%) in the old tests (first use) were classified as poor while there were 37% poor items (10 of 27 items) in the new tests (second use). However, these results are limited because the number of revised items in this course was far fewer than the number of non-revised items (6.67% revised items and 93.33% of non-revised items).

The results might have occured for several reasons. First, the revised and non-revised items were still too difficult for all students, therefore the items still

had poor discrimination (can not discriminate between good and poor students). The revised and non-revised items also may still ambiguous for the students. Further improvement might be obtained by clarifying that 'good' items do not equal hard items. The writers or revisers may need to see a sample of items that might be defined as hard with good discrimination, hard with poor discrimination, moderate with good discrimination, moderate with poor discrimination, easy with good discrimination, and easy with poor discrimination. Second, UT students might be scoring poorly on either the revised or non-revised items because of the possible influence of the course material being difficult (subject matter issue). It is recognized that Mathematics courses, such as Linear Algebra I, are difficult subjects. Conventional university students are able to communicate directly with their professors to meet difficulty in understanding such courses. However, UT students rarely have the option of face-to-face contact.

Appendix D also shows that revisions in FMIPA Statistics program did not provide a lot contribution to changes in individual item discrimination. In the course, Statistical Method I (STAT4110), 1 of 8 revised items (12.5%) showed improvement in discrimination,

while 8 of 22 non-revised items (36%) improved. In the course, Applied Experimental Design (STAT4213), 2 of 6 revised items (33%) showed improvement in discrimination while 10 of 24 non-revised items (42%) improved. In the course, Survey Sampling Method (STAT4334), 2 of 5 revised items (40%) showed improvement while 12 of 25 non-revised items (48%) improved in discrimination. These results were supported by data given in Table 18. Table 18 indicates that the average discrimination of revised items in Statistics program was 0.014 lower in the new test while the average of non-revised items was 0.004 higher. In the course, Statistical Method I (STAT4110), the average discrimination of revised items was 0.044 lower in the new test while that of non-revised items was only 0.023 lower. For Applied Experimental Design (STAT4213), the average discrimination of revised items was 0.002 lower in the new test while the average of non-revised items was 0.002 higher. In the course, Survey Sampling Method (STAT4334), the average discrimination of revised items was only 0.019 higher in the new test while the average of non-revised items was 0.029 higher. These results indicate that, in those courses, the item revisions did not provide desired changes in the average item discrimination. Table 19, however, shows that, in

Applied Experimental Design (STAT4213) and Survey Sampling Method (STAT4334), the revisions resulted in desired changes in the percentages of items judged to be good/very good and poor. For revised items in Applied Experimental Design (STAT4213), the percentage of good/very good items was 33.3% higher in the new test while for non-revised items, it was 8.3% lower. In addition, the percentage of revised items judged to be poor did not change while the percentage of non-revised items was 4.1% higher. For the revised items in Survey Sampling Method (STAT4334), the percentage of good/very good items was 40% higher in the new test while for non-revised items, it was only 20% higher. In this course, the percentage of revised items judged to be poor did not change while the percentage of non-revised items was 8% higher.

For Statistical Method I (STAT4110), on the other hand, the item revisions did not provide desired changes in the percentages of good/very good and poor items (see Table 19). In this course, the percentage of revised items judged to be good/very good was 12.5% lower while the percentage of non-revised items was only 9.1% lower. In addition, the percentage of revised items judged to be poor in this course was 12.5% higher while the percentage of non-revised items did not change. These

results are not surprising since after revisions, the items were still in the very difficult category. It is recognized that an item which is too hard for a group of students may produce a poor discrimination (can not discriminate between good and poor students). Further improvement might be obtained by making the stem more clear and revise the alternatives which still had positive discriminations. The alternatives may still be ambiguous for the good students.

In summary, item revisions in two of three Physics courses in the Faculty of Education (FKIP), Alternating Current (PFIS4439) and Introduction to Quantum Mechanics (PFIS4437), and in two of three English courses, English for Arts and Science (PING4447) and English for Education (PING4441), resulted in desired changes in both the average item discrimination and the percentages of good/very good and poor items. In those courses, the average discrimination of revised items increased (the changes in a range of 0.027 to 0.106) while the averages of non-revised items tended to decrease (the changes in a range of a 0.033 to 0.074). In addition, the percentages of revised items judged to be good/very good in those courses increased (the changes in a range of 11.6% to 24.2%) while the percentages of non-revised items tended to decrease (the changes in a range of

14.9% to 17.4%). The percentage of revised items judged to be poor in those courses decreased (the changes in a range of 13.8% to 27.2%) while the percentages of non-revised items were not reduced (the changes in a range of 0% to 22%).

In the course, Atomic Physics (PFIS4438), on the other hand, the item revision only provided desired changes in the percentage of good/very good and poor items. In this course, the percentage of revised items judged to be good/very good increased while the percentage of items judged to be poor decreased. Meanwhile, the percentage of non-revised items judged to be good decreased and the percentage of non-revised items judged to be poor increased.

In the course, Business English (PING4448), the item revisions did not provide desired changes in either the average item discrimination or the percentage of good/very good and poor revised items. In this course, the average discrimination of both revised and non-revised items decreased. In addition, the percentage of both revised and non-revised items judged to be good/very good decreased while the percentage of both revised and non-revised items judged to be poor increased.

Within the Faculty of Mathematics and Natural

Science (FMIPA), only item revisions in one of two Mathematics courses, Calculus I (MATK4110), resulted in desired changes in both the average item discrimination and the percentage of good/very good and poor items. In this course, the average discrimination of revised items was 0.056 higher in the new test while the average of non-revised items was only 0.038 higher. In addition, the percentage of revised items judged to be good/very good was 15.4% higher in the new test while the percentage of non-revised items was only 5.8% higher. The percentage of revised items judged to be poor in this course was 23% lower while the percentage of non-revised items was only 5.9% lower.

In Linear Algebra I (MATK4112), item revisions did not provide desired changes in either the average item discrimination or the percentage of good/very good and poor items. In this course, the average discrimination of both revised and non-revised items decreased. In addition, the percentage of both revised and non-revised items judged to be good/very good decreased while those items judged to be poor increased. However, these results were limited since the number of revised items in this course (MATK4112) was far fewer than the number of non-revised items.

Of the three Statistics courses, item revisions in

two courses, Applied Experimental Design (STAT4213) and Survey Sampling Method (STAT4334), only provided desired changes in the percentage of good/very good and poor items. In these courses, the percentages of both revised and non-revised items judged to be good/very good were higher in the new test with the greatest changes occurring in the revised items. In addition, the percentages of revised items judged to be poor in these courses did not change while the percentages of non-revised items were higher (the changes were 4.1% for Applied Experimental Design-STAT4213 and 8% for Survey Sampling Method-STAT4334).

Meanwhile, in Statistical Method I (STAT4110), the item revisions did not provide desired changes in either the average item discrimination or the percentage of good/very good and poor items. In this course, the average discrimination of revised items was 0.044 lower in the new test while those of non-revised items was only 0.023 lower. In addition, the percentage of revised items judged to be good/very good in this course was 12.5% lower while the percentage of non-revised items was only 9.1% lower. The percentage of revised items judged to be poor in this course was 12.5% higher while the percentage of non-revised items did not change.

Reliability

In order to determine whether the quality of test reliabilities between faculties was different and to judge whether test reliabilities in both faculties changed, both Ebel's and UT's criteria for test reliability were applied to the old (87.1, 88.1, and 89.1) and new (92.1) tests. Differences in test reliabilities were examined through the analysis of the individual test reliability coefficient and the number of tests which meet the criteria for reliability. Faculty-by-faculty, program-by-program, and course-by-course comparisons were conducted.

As shown in Tables 20 and 21, across the sample of three test administrations, two of six tests (33.3%) offered in test administrations 87.1 and 92.1 in the Faculty of Education (FKIP) met reliability standards recommended by Ebel (1972), and three of the six tests (50%) provided in test administration 89.1 also met these standards. Those tests are the tests of English courses. No tests of Physics courses met reliability standards recommended by Ebel (see Table 21). In the English program, all tests (3 of 3 tests) offered in every test administration met UT's standards in reliability with the number of tests which had strong reliability being higher than the number of tests which

Table 20
Distribution of Reliability Coefficient Between Faculties

Reliability coefficient range	Number of tests in each category					
	FKIP			FMIPA		
	87.1	89.1	92.1	88.1	89.1	92.1
Ebel's criteria						
Desired values:						
≥ 0.80	2	3	2	0	0	0
UT's criteria						
Desired values:						
≥ 0.60	4	5	5	2	2	2
< 0.60 (weak)	2	1	1	3	3	3
0.60 - 0.80 (moderate)	2	2	3	2	2	2
> 0.80 (strong)	2	3	2	0	0	0

Table 21
Distribution of Reliability Coefficient within FKIP

Reliability coefficient range	Number of tests in each category					
	Physics			English		
	87.1	89.1	92.1	87.1	89.1	92.1
Ebel's criteria						
Desired values:						
≥ 0.80	0	0	0	2	3	2
UT's criteria						
Desired values:						
≥ 0.60	1	2	2	3	3	3
< 0.60 (weak)	2	1	1	0	0	0
0.60 - 0.80 (moderate)	1	2	2	1	0	1
> 0.80 (strong)	0	0	0	2	3	2

had moderate reliability (see Table 21). Within the Physics program, on the other hand, one of the three tests (33.3%) offered in test administration 87.1 and two of the three tests (66.7%) provided in test administrations 89.1 and 92.1 met UT's standard with all of these being in the moderate category (see Table 21). The tests were tests 89.1 and 92.1 of Alternating Current (PFIS4439), test 89.1 of Atomic Physics (PFIS4438), and tests 87.1 and 92.1 of Introduction to Quantum Mechanics (PFIS4437) (see Table 5).

In the Faculty of Mathematics and Natural Science (FMIPA), on the other hand, none of the tests used in this study met reliability standards as recommended by Ebel (see Table 20). However, two of the five tests (40%) provided in test administrations 88.1, 89.1, and 92.1 met UT's standards and all of them were in the moderate category. The tests were tests 88.1 and 92.1 of Calculus I (MATK4110), tests 88.1 and 89.1 of Linear Algebra I (MATK4112), and tests 89.1 and 92.1 of Survey Sampling Method (STAT4334). It appears that the number of tests in the Mathematics program which meet UT's standards was higher than the number in the Statistics program.

Furthermore, based on a comparison of old and new tests, the new tests in two of the six courses in the Faculty of Education (FKIP), Alternating Current (PFIS4439) and Introduction to Quantum Mechanics (PFIS4437), had a higher reliability (see Table 5). However, the new reliability coefficient did not meet Ebel's standards (still in the moderate category). These results are not surprising, since in these courses, the percentage of items judged to have acceptable difficulty under UT's criteria was higher while the percentage of poor items (items with discrimination less than 0.20) was lower in the new test

(see Tables 7 and 14). In the courses, English for Arts and Science (PING4447) and Business English (PING4448), on the other hand, the reliability coefficients of the new tests were lower than the reliability of the old tests 87.1 and 89.1. Meanwhile, reliability of the new tests of Atomic Physics (PFIS4438) and English for Education (PING4441) were lower than the reliability of the old test 89.1 but slightly higher than the reliability of the old test 87.1. These results are also not surprising because the new tests of these courses did not have a higher percentage of items judged to have acceptable difficulty under both Ebel's and UT's criteria (see Table 7). In addition, the new tests of these courses did not have a lower percentage of poor items (discrimination < 0.20) or a higher percentage of good/very good items (discrimination $\geq$ 0.30) (see Table 14).

In the Faculty of Mathematics and Natural Science (FMIPA), the new tests in two of the five courses, Applied Experimental Design (STAT4213) and Survey Sampling Method (STAT4334), had a higher reliability coefficient than the old tests 88.1 and 89.1 (see Table 5). However, the new reliability coefficient did not meet Ebel's standards (still in the weak and moderate category). These results are not surprising because the

Table 22
Distribution of Reliability Coefficient within FMIPA

Reliability coefficient range	Number of tests in each category: Mathematics 88.1	Mathematics 89.1	Mathematics 92.1	Statistics 88.1	Statistics 89.1	Statistics 92.1
Ebel's criteria						
Desired values:						
≥ 0.80	0	0	0	0	0	0
UT's criteria						
Desired values:						
≥ 0.60	2	1	1	0	1	1
< 0.60 (weak)	0	1	1	3	2	2
0.60 - 0.80 (moderate)	2	1	1	0	1	1
> 0.80 (strong)	0	0	0	0	0	0

new tests of these courses had a higher percentage of items judged to have acceptable difficulty and had a higher percentage of very good items (discrimination > 0.399) (see Tables 8 and 15). In Linear Algebra I (MATK4112), on the other hand, the reliability coefficient of the new test 92.1 was lower than the reliability of the old tests 88.1 and 89.1. This result is not surprising since the new test of this course had a lower percentage of items judged to have acceptable

difficulty, a higher percentage of poor items (discrimination < 0.20), and a lower percentage of very good items (discrimination > 0.399) (see Tables 8 and 15). Meanwhile, the reliability coefficients in the new tests of Calculus I (MATK4110) and Statistical Method I (STAT4110) were lower than the reliability of the old test 87.1 but higher than the reliability of the old test 89.1. The results in the both faculties could be affected by the nature of both examinees and tests. The homogeneity of the examinees in ability and the homogeneity of the items in the area being measured, the time limit, the test length, and the number of items which have high discrimination and moderate difficulty may affect the reliability coefficient (Crocker & Algina, 1986; Ebel, 1972). Data provided in Tables 23 and 24 support Crocker's and Algina's, and Ebel's opinions. It appears from Tables 23 and 24, that there were positive correlations between the reliability coefficients and the number of items, the number of good items (items with discrimination indices equal to or higher than 0.30 and item difficulty in a range of 0.30 to 0.70), and the average raw scores of tests used in this study. There were also positive correlations between the number of items and the number of good items, between the number of items and average raw

Table 23
Relation of The Number of Items, Number of Good Items, and Average Raw Score to Reliability Coefficient

Course	Period	No of items	No of Good items (Items with Discr $\geq$ 0.30 and 0.30 < p < 0.70)	Average raw score	Reliability coefficient
STAT4334	88.1	25	0 (0.00%)	8.29 (33.16%)	-0.056
STAT4213	89.1	30	6 (20.00%)	9.91 (33.03%)	0.265
STAT4110	89.1	30	6 (20.00%)	9.55 (31.83%)	0.285
STAT4110	92.1	30	4 (13.33%)	11.50 (38.33%)	0.326
STAT4213	88.1	30	5 (16.67%)	9.58 (31.93%)	0.389
STAT4213	92.1	30	6 (20.00%)	11.59 (38.63%)	0.420
STAT4110	88.1	29	6 (20.00%)	10.22 (35.24%)	0.421
PFIS4438	87.1	44	9 (20.45%)	15.13 (34.39%)	0.427
MATK4110	89.1	30	8 (26.67%)	10.76 (35.87%)	0.433
PFIS4438	92.1	50	7 (14.00%)	20.79 (41.58%)	0.464
MATK4112	92.1	30	7 (23.33%)	9.98 (33.27%)	0.475
PFIS4437	89.1	50	6 (12.00%)	17.45 (34.90%)	0.511
PFIS4439	87.1	40	15 (37.50%)	16.89 (42.23%)	0.586
STAT4334	89.1	27	7 (25.93%)	10.45 (38.70%)	0.642
PFIS4437	87.1	45	16 (35.56%)	18.08 (40.18%)	0.656
MATK4110	92.1	30	9 (30.00%)	12.12 (40.40%)	0.661
MATK4112	89.1	28	11 (39.29%)	10.18 (36.36%)	0.667
PFIS4437	92.1	49	12 (24.49%)	19.61 (40.02%)	0.685
STAT4334	92.1	30	13 (43.33%)	11.83 (39.43%)	0.692
PFIS4438	89.1	60	19 (31.67%)	25.45 (42.42%)	0.696
PFIS4439	89.1	50	13 (26.00%)	25.82 (51.64%)	0.702
MATK4110	88.1	30	12 (40.00%)	12.40 (41.33%)	0.719
PFIS4439	92.1	49	15 (30.61%)	23.52 (48.00%)	0.727
MATK4112	88.1	24	9 (37.50%)	9.19 (38.29%)	0.750
PING4441	87.1	67	10 (14.93%)	37.06 (55.31%)	0.772
PING4447	92.1	70	14 (20.00%)	43.25 (61.79%)	0.796
PING4448	92.1	60	13 (21.67%)	39.42 (65.70%)	0.802
PING4447	89.1	80	12 (15.00%)	52.04 (65.05%)	0.810
PING4447	87.1	70	23 (32.86%)	36.91 (52.73%)	0.812
PING4441	92.1	60	23 (38.33%)	35.31 (58.85%)	0.849
PING4448	87.1	70	16 (22.86%)	43.31 (61.87%)	0.855
PING4448	89.1	60	27 (45.00%)	35.55 (59.25%)	0.863
PING4441	89.1	80	38 (47.50%)	41.96 (52.45%)	0.894

Table 24
Correlation Matrix

	Reliability coefficient	No of items	No of Good items	Average raw score
Reliability coefficient	1			
No of items	0.67**	1		
No of Good items	0.75**	0.69**	1	
Average raw score	0.71**	0.97**	0.65**	1

Note. ** significant at $p < .001$
under two-tailed significancy test

score, and between the number of good items and average raw score. The corelation between the number of items and average raw score was strong (r=0.97), while the correlations for the others were moderate (r in a range of 0.65 to 0.75). Under the two tailed significancy test, the correlations were significant at $p < .001$. The results indicate that the reliability coefficient increased as the number of items, the number of good items, and the average raw score increased. Meanwhile, the number of good items and the average raw score increased as the number of items increased. In addition, the average raw score increased as the number of good items increased.

In summary, across the sample of three test administrations, some of the English tests in the

Faculty of Education (FKIP) used in this study met reliability standards recommended by Ebel, while no test in the Faculty of Mathematics and Natural Science (FMIPA) met the standards. However, in addressing UT's standards, two of the five FMIPA tests provided in each test administration met those standards. Two of the five courses in FMIPA (Statistics courses) had a higher reliability coefficient in the new test 92.1, while two of six courses in FKIP (Physics courses) did.
However, the new reliability coefficients were still in the weak and moderate category. These results could be affected by the nature of both examinees and items.

CHAPTER V

Discussion, Conclusions, Limitations, Implications, and Recommendations

Discussion

Ignoring, for a moment, the limitations of the study, there were three important findings: (1) most of the tests, both before and after revision, in the Faculty of Education (FKIP) and the Faculty of Mathematics and Natural Science (FMIPA) used in this study did not meet Ebel's criteria in item difficulty, item discrimination, and test reliability; (2) item revisions in some FKIP and FMIPA courses resulted in desired changes in item difficulties, item discriminations, and test reliabilities; (3) FKIP tests still have 50 to 90% revised items and 30 to 80% non-revised items with discrimination less than 0.30 while FMIPA tests have 60 to 100% revised items and 40 to 80% non-revised items with discrimination less than 0.30, implying that further revision should be conducted in both FKIP and FMIPA test items.

More conclusively, most of the tests, both before and after revisions, in both faculties FMIPA and FKIP used in this study were much harder, less discriminating, and less reliable than norm-referenced test (NRT) experts posit as ideal, especially for

science courses such Physics, Mathematics, and Statistics. However, considering UT's standards, the results show that the percentage of items with UT's acceptable values in item difficulty and discrimination was high enough (about 80-90% items for item difficulty and about 60-70% items for item discrimination) with most of the items in the difficult/very difficult category and in the marginal/ reasonably good discrimination. In addition, across the three test administrations, 40% of tests in FMIPA and more than 60% of tests in FKIP met UT's standards of reliability. Thus, it would seem UT can maintain its standards in item difficulty and discrimination level because most of its items meet these standards. Although UT's standards are different than those recommended by Ebel (testing expert), the standards may be appropriate for UT which provides the first distance learning in Indonesia. The nature of UT students who have different characteristics from other Indonesian university and also from other distance-education university students may allow UT to use different standards than those recommended by testing experts. In comparison to other Indonesian university students, UT students are more heterogeneous in ability levels, in ages, and in social-economical status, because UT's policies enable high

school graduates (including vocational school graduates) and working people throughout Indonesia, to study at UT without any entrance examination. Those who took social-study program during high school, for example, can take Mathematics or Statistics program at UT without any entrance examination. In addition, unlike distance-education university students in western countries who are used to study independently and autonomously, UT students are not ready to become active, autonomous, and persistent in isolated learning situations. This lack of preparation is historical. From elementary through high school, they perceived learning as a relationship with a teacher, which is immediate, oral, and hierarchical. For Indonesian learners, learning is a passive activity but teaching is an active one. The teacher is an authority figure, commanding automatic, and unchallengeable respect while learners are obligated to defer to the knowledge and authority of the teacher and must seek approval, direction, and permission at every step of their passage through education (Dunbar, 1991).

Furthermore, since most items, especially for science courses, were hard and classified as marginal and reasonably good; and most of the tests had low means of difficulty (less than 0.40) and low means discrimination

(in the marginal category), important points are suggested. First, the need to study the difficulty of mastering the written materials (modules) of the science courses used in this study and to investigate the consistency of their contents and the item content. The students may feel the tests are too difficult because the modules lack explanations and contain errors which hinder their understanding of the modules. The students may also feel the tests are too difficult when the test item content is not discussed in the modules. Second, UT's test writers, especially those who develop tests for science courses, should reconsider their test items. It seems that UT's test writers must reconsider whether the difficulty level of an item is appropriate to the level of ability and range of understanding of UT's students. The test writers' awareness of the students' level of ability is necessary because, when a test is too hard or too easy for a group of students, restriction of score range and of true score variance is likely to be the result; consequently, the internal consistency coefficient will be low (Ebel, 1972). Lack of awareness on the part of UT's test writers about the range of UT students' ability and understanding is likely to be the reason why most of the items are too difficult and, consequently, why the items are less

discriminating and less reliable. The lack of awareness perhap occurs because of the nature of UT and the status of its test writers. Most of UT's test writers are not full time staff of UT. The writers are professors from traditional top state universities who usually teach and prepare tests for students who have higher levels of ability than UT's students and who possess different characteristics from UT's students. The students at their own institutions are qualified high school graduates who passed the national entrance examinations, while the students at UT are high school graduates who are accepted as UT students without any entrance examinations. In addition, unlike in their own institutions where they are able to communicate directly with their students, the test writers, who are usually also the course writers, are not required to engage in direct contact with UT students and are not involved in any of the feedback processes, such as involvement in occasional in-person lectures and seminars, and in assignment and examination marking. To the UT course writers and test writers, their students are disembodied and anonymous. It was suggested here to incorporate more direct lines of communication between course writers or test writers and students including involvement in learning processes, such as face-to-face and written

tutorials, and seminars; and in test result evaluations. A more direct line of communication between UT test writers and students may develop the awareness of the test writers as to the range of UT students' ability. The difference between test formats which the test writers develop at UT and at their own institutions is likely to be another factor accounting for the low item discrimination. At their own institutions, UT's test writers usually develop essay items while at UT they must develop 20 to 90 multiple-choice items for a 90-minute test. These differences demand that UT's test writers have writing skills for multiple choice tests which are different from those required for essay tests. In developing multiple choice items, the writers are required to have the ability to make all alternatives plausible and attractive to both the less knowledgable and the less skillful students; whereas in developing essay items, these abilities are not required. The need for test writers to spend time working on UT's concerns seems to be another factor required for improving the quality of UT's test items. The writers must be willing to spend sufficient time and energy to do a competent job because developing multiple choice items require an extensive amount of time in writing and revising.

In terms of test reliability, since the percentage of

tests studied which meet UT's standards was only 60.6, UT should determine how these standards can be met through better test preparation by its test writers. Since poor internal reliability arises from poor sampling of one source of content of the test (Crocker & Algina, 1986; Wood, 1961), it seems that UT should ask its test writers to reconsider the Table of Specifications in order to improve their internal reliability coefficient. The tests developed by UT's writers could be drawn from too small areas of the Table of Specifications. Even if it is large, it might not be representative since the Table of Specifications over-emphasizes certain areas and neglects others. If the Table of Specifications were representative, UT's test writers could construct more reliable tests by controlling the factors affecting reliability. Ebel (1972) argued that the reliability coefficient will be greater for scores from a longer test than from a shorter test, from a test composed of more homogeneous items than from a more heterogeneous test, from a test composed of more discriminating items than from a test composed of less discriminating items, from a test whose items are of middle difficulty than from a test composed from mainly quite difficult or quite easy items, from a group having a wide range of ability than from a group

more homogeneous in ability, from a speeded test than from one all examinees can complete in the time available. This means UT's test writers can improve the test reliability through writing, revising, and selecting test items of high discrimination and moderate difficulty. This also means that UT's test writers should include as many items as possible in the test, so as to make the test as long as possible. With the limited time for the test (90 minutes), UT's test writers (especially those who develop tests for science courses) can favor items that are least time-consuming individually in order to produce a longer test. However, if the cost required to increase UT's test reliability proves prohibitive, UT may use the most reliable of the available tests even if they have a reliability of only 0.40 or 0.50. A test with low reliability may still have some validity and can therefore be useful (Mehrens & Lehmann, 1973). Ideally, of course, it would be better to develop or select a test having higher rather than lower reliability, but in some situations where all tests may yield equally low reliability or the cost required to increase reliability may be prohibitive, teachers may use the most reliable of the available tests even if they have a reliability of only 0.40 or 0.50 (Sax, 1974;

Thorndike, 1951). Furthermore, although some of the tests included in this investigation met the difficulty, discrimination, and reliability standards recommended by either Ebel or UT, to establish definitively that these tests are beyond criticism would require additional analyses of the language used in the questions and an examination of the course content and structure. Since data for these analyses were not available, the author of this study made no assumptions about other important issues such as fairness, representativeness, and content validity.

In detail, the second key finding was that item revisions in one of three FKIP Physics courses used in this study, Introduction to Quantum Mechanics (PFIS4437) and in two of three English courses, English for Arts and Science (PING4447) and English for Education (PING4441), resulted in desired changes in both item difficulty and discrimination. Meanwhile, item revisions in Atomic Physics (PFIS4438) only provided desired changes in item difficulty and item revisions in Alternating Current (PFIS4439) only provided desired changes in item discrimination. These findings support Lange, Lehmann, and Mehrens' findings (1967) that poor items (items with low discriminations) could be improved through revision using a complete item analysis.

Although improvement in discrimination of individual revised items was found in most of FKIP courses used in this study, the results can not be generalized to all courses in FKIP. Further investigation will be required to determine whether item revision in FKIP courses improves its individual item discrimination. It is also difficult to conclude that the whole item revision process in FKIP (the process of deciding whether or not an item will be revised) was successful since some of the new tests, consisting of both revised and non-revised items which were selected based on the results of revision activity, were not better in item discrimination and test reliability than those in the old tests consisting of untested items. Failure to find a more reliable test could occur due to several factors. The first factor is the nature of items. At UT, a multiple-choice test may consist of five types of items. The changes in the composition of the number of each type of items; in the arrangement of the items including the order of the topics, cognitive domains, item format, and item difficulties; in the homogeneity of items being measured; and in the number of items may contribute to a decrease in item discrimination and test reliability. In addition, this study found that there were positive correlations between test reliabilities and the number

of items, the number of good items (items with discrimination indices equal to or higher than 0.30 and difficulty in a range of 0.30 to 0.70), and the average scores of tests. This means that test reliability increases as the number of items, the number of good items, and the average score increase. A second factor is the nature of students. The differences in the range of abilities of students who took the new and old tests may contribute to the decrease in item discrimination and test reliability. Ebel (1972) argued that the reliability coefficient will be greater for scores from a group having a wide range of abilities than from a group more homogeneous in abilities. In this case, FKIP students who took the new tests 92.1 may have a more homogeneous ability level than the students in the old tests 87.1 and 89.1. The third factor is the nature of writers and revisers. Some revisers and writers in FKIP do not believe in statistical matters such as item difficulty and discrimination indices in item analysis and, consequently, they refuse to revise the hard items with low discrimination indices and they accept the items as non-revised items. Disbelief of item analyses on the part of FKIP test writers may occur due to their unfamiliarity with that method. At their own institutions, FKIP test writers provide essay items to

their small number of students and do not use item analysis in rechecking the effectiveness of their tests. At UT, the test writers are also not involved in the item-analysis process which is conducted after test administration. The item-analysis method was just introduced to the writers when UT asked the writers to revise the items in 1989. Since UT's establishment in 1984, UT has conducted one item revision (once in five years). It seems that the awareness of FKIP's test writers about item-analysis method and their understanding about the importance of that method should be improved. It also seems that test writers in FKIP-UT should not rely too much on their own subjective judgments about the quality of the items but should pay more attention to the item statistics in item analysis in order to produce a more reliable test. Although item statistics do not ensure the quality of the items since the values can fluctuate when different samples of students are tested, the item statistics can yield useful information regarding the success of a writer's test. The item difficulty indices, for example, can be used to examine whether a test writer's judgments about the difficulty level of the items was too high for the level of ability and range of understanding of FKIP students. In discussing such issues, Mehrens and

Lehmann (1978) recommended the use of some quantitative evidence such as item-analysis data to support an item writer's judgments about the item difficulty level. The recommendation was offered in regards to the evidence demonstrating that there were many professional item writers and classroom teachers who misjudged the difficulty of their items. Popham (1981) also suggested the use of both judgmental and empirical (statistical) techniques in improving test items with the use of empirical strategies preceding the application of judgmental strategies.

Furthermore, the results also indicated that the average discrimination of both revised and non-revised items in the new tests tended to remain in the marginal category with the percentage of poor items (discrimination less than 0.20) still relatively high (most of tests had 30 to 40% revised items and 20 to 30% non-revised items which were still categorized as poor). The percentages of items with discrimination less than 0.30 were also still high. Most of the tests had 50 to 90% revised items and 30 to 80% non-revised items with discriminations less than 0.30. Since UT's criteria considers items with discrimination less than 0.30 in need of revision, these outcomes suggest that further revision should be conducted at FKIP. Further

improvement might be obtained by clarifying how the writers are to change the items rather than give instruction to simply shift the item difficulty and discrimination. For example, the writers may need to see a sample of items that might be defined as easy, moderate, and hard; and that have poor, marginal, good, and very good discriminations.

In FMIPA, item revisions in one of the two Mathematics courses (Calculus I-MATK4110) and in one of the three Statistics courses (Survey Sampling Method-STAT4334) resulted in desired changes in both item difficulty and discrimination since they improved the average item difficulties and discriminations of their revised items, increased the percentage of good items, and reduced the percentage of poor items. However, it is difficult to conclude whether or not item revision in other courses used in this study was effective because the number of revised items included in the tests was far fewer compared to the number of non-revised items. Conclusively, it is difficult to say that revision in FMIPA tests was not successful in improving individual item discrimination. As in FKIP, it is also difficult to conclude that the whole item revision process in FMIPA was successful since some of the new tests composed of both revised and non-revised items which

were selected based on the results of revision activity were not better in item discrimination and test reliability than those in the old tests.

Factors affecting these outcomes might be the same as those in FKIP. As in FKIP, the new average item discrimination of both revised and non-revised items of the tests was likely to remain in the marginal category with the percentage of poor items still being high (most of tests had 60 to 70% revised items and 20 to 30% non-revised items which were still classified as poor). The percentages of items with discrimination less than 0.30 were also still high. Most of tests had 60 to 100% revised items and 40 to 80% non-revised items with discriminations less than 0.30. Given UT's criteria which suggest that items with discrimination less than 0.30 should be revised, these outcomes suggest that further revision should also be conducted at FMIPA.

In summary, most of the science tests, both before and after revisions, in the Faculty of Education (FKIP) and the Faculty of Mathematics and Natural Science (FMIPA) used in this study were in the difficult/very difficult category, with marginal discrimination and weak/moderate reliability. These outcomes suggest two points. First, the need to study the difficulty of mastering the written materials (modules) of the science

courses used in this study. If, it is found that the students have difficulty in understanding the modules, UT should revise the modules in a way that makes students easier to learn. Collaboration between instructional-design experts and module writers (subject-matter experts) in revising the modules may yield far better modules than would be produced by either module writers or instructional-design experts alone. Second, the skills in writing and revising items should be improved. However, the process of improving those skills is not easy. It is recognized that the process of constructing good test items and of revising poor test items is not simple. Not all individuals are equipped to master writing and revising skills easily. The abilities required, such as skills in written expression or verbal communication and abilities to understand the ability level of students, can not be produced in a short period of time because these abilities grow slowly. Manuals and rules may provide useful guides and helpful suggestions for item writing and revising, but there is no automatic process for the construction of good test items. Even though the item writers or revisers might possess the required abilities, their success in writing good test items or in revising poor items could vary depending on the

amounts of energy and willingness the writers will devote to the task. The opportunity UT offers to the test writers to be more involved in learning processes and in test result evaluation, and their willingness to spend more time on UT's concerns are one the important factors required in improving the quality of UT's test items.

Conclusions

Based on the results of descriptive data on the test results of six courses in the Faculty of Education (FKIP) and five courses in the Faculty of Mathematics and Natural Science (FMIPA) drawn from three test administrations, it is concluded that most of the test items in FKIP and FMIPA have met UT's standards. Overall, FKIP and FMIPA have the same approximate qualities in item difficulty. However, the qualities of item discrimination and test reliability were slightly higher in FKIP. Most of the items in both faculties are still in the difficult/very difficult category, with marginal discrimination. Most of the tests in both faculties are also harder, less discriminating, and less reliable than testing experts posit as ideal, especially for science courses such as Physics, Mathematics, and Statistics.

The other major conclusion of this study is that

item revisions in one of three Physics courses in FKIP used in this study, Introduction to Quantum Mechanics (PFIS4437), and in two of three English courses, English for Arts and Science (PING4447) and English for Education (PING4441) resulted in desired changes in both item difficulty and discrimination. Meanwhile, item revisions in Atomic Physics (PFIS4438) only provided desired changes in the item difficulty and item revisions in Alternating Current (PFIS4439) only provided desired changes in item discrimination. However, only item revisions in English for Education (PING4441) resulted in significant differences in the average item discrimination. In considering a set of revised and non-revised items in the new tests, however, only the new tests of two Physics courses (Introduction to Quantum Mechanics-PFIS4437 and Alternating Current-PFIS4439) have higher means of item discrimination and test reliability than the two old tests. In terms of item difficulties, only one of the three Physics courses, Introduction to Quantum Mechanics (PFIS4437), had a higher percentage of acceptable items and moderate items, and a lower percentage of items judged to be difficult or very difficult ($p \leq 0.40$) in the new test. In the Faculty of Mathematics and Natural Science (FMIPA), it is difficult to conclude that item revisions

resulted in desired changes in item difficulty and discrimination of a test since the number of revised items in the new test is far smaller than that of non-revised items. However, regardless of the small number of revised items, item revision in one of the two Mathematics courses (Calculus I-MATK4110) and in one of the three Statistics courses (Survey Sampling Method-STAT4334) resulted in desired changes in item difficulty. However, the item revision in these courses did not appear to increase the percentage of items which have acceptable difficulty. Item revisions in Calculus I (MATK4110) also resulted in desired changes in the average item discrimination and percentage of good and poor items, while item revision in Survey Sampling Method (STAT4334) only provided desired changes in the percentage of good and poor items. Item revisions in Applied Experimental Design (STAT4213) also only provided desired changes in the percentage of good and poor items. Regarding a set of revised and non-revised items in the new test, only the new tests of two Statistics courses, Applied Experimental Design (STAT4213) and Survey Sampling Method (STAT4334), had a higher test reliability coefficient than the old tests, and none had a better mean of item discrimination. Only the distribution of item difficulties in the Statistics

tests changed a lot so as to have any impact on the measurement of student performance.

Overall, since the percentages of revised and non-revised items with discrimination less than 0.30 were still in the 50-90% and 30-80% range respectively for the Faculty of Education (FKIP), and were in the 60-100% and 40-80% range respectively for the Faculty of Mathematics and Natural Science (FMIPA), it could be concluded that further revision should be conducted in both FKIP and FMIPA test items.

Limitations

These conclusions should be considered in light of the following limitations. One limitation is that the revised and non-revised items in each of two faculties were drawn from a small sample size of test administrations and courses. Therefore, the generalizability of the effects of the revisions on item characteristics at each faculty is limited.

The second limitation is that the number of revised items in FMIPA's new tests was far smaller than the number of non-revised items. As a result, it is difficult to generalize the effectiveness of item revision in a FMIPA course.

Another limitation, the use of item analyses after deletion of some items, resulted from the problem of

finding the item analyses before deletion. However, since the percentage of eliminated items was relatively small (less than or equal to 10%), the effect of item elimination on item characteristics was slight.

The final limitation comes from some problems with item analyses. Some authors (Huck & Bowers, 1972; Lien, 1976; Tinari, 1979) provided possible factors which influence the item statistics of item analyses. The possible factors are: (1) the examinee guessing, (2) the conditions under which the test is given, (3) the location of the correct answer among the response alternatives, and (4) the serial location of the item within the test.

Implications

A major implication of these findings is that UT's test writers appear to need better training in preparation, analysis, and revision of multiple-choice test items. This could take the form of a laboratory approach where in (1) the participant is actively involved in solving a problem, (2) the problem situation is simulated as realistically as possible, (3) quantifiable data are produced and recorded to reveal the nature of the response of the participants, (4) feedback on data is provided to permit each participant to contrast his reactions with those of the others, (5)

data are discussed and analyzed so as to lead to generalizations and implications for practice (Harris, Bessent, & McIntyre, 1969).

The training activities could include (1) discussing the importance of formal testing in the evaluation of instruction; (2) demonstrating the ways in which teacher-made tests reflect and influence students' learning; (3) demonstrating some of the important characteristics of good tests and good test items (validity, reliability, difficulty, and discrimination); (4) discussing some basic problems in test construction, such as inconsequential content being tested, learning being tested only at superficial recognition or recall levels, test items constructed so as to be nondiscriminating or negatively discriminating: non-functioning distractors converting multiple-choice items to simple choice items, trick wording misleading students who know, a word used being a specific determiner of the appropriate answer, items tending to be too easy or too difficult; (5) developing skills in identifying weaknesses in teacher-made test items; (6) developing skills needed to improve teacher-made test items.

The training activities could begin with an introductory statement concerning the importance of teacher-made

tests, the general characteristics of good tests and test items, and some basic problems in test construction. Several collections of items made by several UT's test writers could be presented to the participants in the form of a simulated test. The items would then be discussed, and their strengths and shortcomings revealed. Several ways of evaluating teacher-made tests, such as the use of item analysis could then be presented. Further, the participants could be asked to revise the poor items and discuss the revised items. At the end of the training session, the participants could be asked to construct a set of items which will be administered to the students. Follow up studies of the test constructed after the training may be required to reveal the types of problems requiring further attention.

A second major implication is the need for systematic evaluation of the results of UT's tests. In order to continue both to improve UT test writers' skills in testing and to improve the quality of the tests, UT should provide opportunities for its test writers to systematically analyze the results from their tests, to compare the findings of these analyses with UT's or testing experts' standards of test quality, and then to revise the items which do not meet the

standards. The feedback from systematically analyzing the revised and non-revised items may contribute to the improvement of the abilities or skills required for constructing good test items and in revising poor items. A one-shot evaluation of test results (once in five years) conducted at the present time by UT's test writers may not improve their writing skills. It was suggested by Ebel (1972) that there is no better way for a teacher or professor to continue to improve his skills in testing and the quality of tests he uses, than to systematicaly analyze the results from his tests and to compare the findings of these analyses with ideal standards of test quality. It was also suggested by Mehrens and Lehmann (1978) that a continual revising and rechecking process leads to increase skill in test construction in that a teacher gradually learns what methods of wording and what type of distractors will work best.

A third major implication is the need for assessing the curricular, instructional, and content validity of UT test items. Since validity of a test determines the quality of a good test, in order to improve the quality of its test items, UT needs to reconsider the extent to which its items are relevant to the objectives stated in its curriculum (curricular validity), to the objectives

stated in written course materials (instructional validity), and to the objectives stated in Table of Specifications (content validity). These activities may entail the following steps: (1) reconsidering the objectives stated in the curriculum, written course materials, and Table of Specifications; (2) selecting a panel of qualified content experts; (3) providing a structured framework for the process of matching items to the objectives stated in the curriculum, written course materials, and Table of Specifications; (4) collecting and summarizing the data from the matching process (Crocker & Algina, 1986).

A fourth implication is the need to look carefully at the appropriateness of the criteria for item difficulty, item discrimination, and test reliability at UT. At the present time, the criteria used at UT is based upon the criteria of norm-referenced tests (NRT). Are these criteria appropriate for UT? This question is important for several reasons. First, the variability of UT student test scores is low. Most experts in test construction (Ebel, 1972; Payne, 1974) say that for a narrow range of test scores, it is more appropriate to use the criteria of criterion-referenced tests (CRT) rather than the criteria of norm-referenced tests (NRT) because the discrimination indices in NRT criteria are

easily affected by the range of test scores. Second, UT does not actually use the statistics data listed in item analysis as the basis for item selection or item assembly. In selecting or assembling the items, UT selects the items by matching the objectives and cognitive domains of the items with those stated in the Table of Specifications. Considering this situation, it seem that UT should decide whether NRT or CRT criteria will be used. If UT prefers to use NRT criteria, UT should use item analysis data as the basis for item selection or assembly. On the other hand, if UT wants to use CRT criteria, UT should revise the Table of Specifications in a way that makes the objectives stated in this tool more specific for use with CRT test development.

Recommendations for Future Research

It would be desirable, in the future, to have additional studies in the area of revised and non-revised items. In particular, the use of a larger sample of courses and test administrations, the use of a better composition of revised and non-revised items in a test, and the use of item analyses before deletion of some items would be suggested. In addition, since the difficulty and discrimination of the alternatives are also important in order to know how well the

alternatives work in distracting the poor students, it would be desirable to also evaluate the difficulty and discrimination of the alternatives of both the original and revised items and then analyze the effect of item revision on those item characteristics.

Since the values of item statistics, such as item difficulty and discrimination, can fluctuate when different samples of examinees are tested, any test that is the result of item selection based on revision activity should be subjected to cross validation: a rechecking of the statistical properties of the new test based on a new sample of examinees (Allen & Yen, 1979; Crocker & Algina, 1986). Therefore, the second recommendation is to conduct another evaluation of the selected revised and non-revised items based on a new sample of examinees in order to see if the same items will function effectively the second time.

It was suggested by Popham (1981) that examinee judgment can be used in improving test items. According to Popham, since examinees have experienced test items in a most meaningful context, examinee judgments can provide useful insight regarding particular items. Therefore, the third recommendation is to research UT examinee perceptions about the quality of UT's tests composed of revised and non-revised items, such as the

difficulty, the clarity of the item structure and options, the consistency of item content and the written course material (module), and the representativeness of the test to the content of the module. Although UT should not fully rely on the examinees' responses (as Popham suggested) the feedback from the examinees could result in better revision of the Table of Specifications and items, and consequently, the quality of the tests will improve. A test evaluation using examinee judgment has not been conducted at UT as of yet.

Since difficulty of a test may be affected by difficulty in understanding the instructional materials (modules), then the fourth recommendation is, to explore UT student perceptions about the quality of modules for the courses used in this study.

BIBLIOGRAPHIES

Adkins, D.C. (1974). Test Construction: Development and Interpretation of Achievement Tests (2nd ed.). Columbus, Ohio: C.E. Merril.

Allen, M.J., & Yen, W.M. (1979). Introduction to Measurement Theory. Monterey, California: Brooks/Cole.

Allison, D.E. (1984). The Effect of Item-Difficulty Sequence, Intelligence, and Sex on Test Performance, Reliability, and Item Difficulty and Discrimination. Measurement and Evaluation in Guidance, 16(4), 211 - 217.

Balch, W.R. (1989). Item Order Affects Performance on Multiple-Choice Exams. Teaching of Psychology, 16(2), 75 - 77.

Beuchert, A.K., & Mendoza, J.L. (1979). A Monte Carlo Comparison of Ten Item Discrimination Indices. Journal of Educational Measurement, 16, 109 - 118.

Bloom, B.S. (Ed.). (1956). Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain. New York: David McKay Company.

Bresnock, A.E. (1989). Multiple-Choice testing: Question and Response Position. Journal of Economic Education, 20(3), 239 - 245.

Brown, F.G. (1981). Measuring Classroom Achievement. New York: Holt, Rinehart and Winston.

Chissom, B., & Chukabarah, P.C. (1985). An Investigation of the Relationship between Item Arrangement and Test Performance. Report Research no. 143. Alabama, U.S. (ERIC Document Reproduction Service no. ED 265 185).

Crocker, L., & Algina, J. (1986). Introduction to Classical and Modern Test Theory. New York: Holt, Renehart and Winston.

Cureton, E.E. (1950). Validity, Reliability, and Baloney. Educational and Psychological Measurement, 10, 94 - 96.

Davis, F.B. (1952). Item Analysis in Relation to Educational and Psychological Testing. Psychological Bulletin, 49, 97 - 121.

Dudycha, A.L., & Carpenter, J.B. (1973). Effects of Item Format on Item Discrimination and Difficulty. Journal of Applied Psychology, 58(1), 116

Dunbar, R. (1991). Adapting Distance Education for Indonesians: Problems with Learner Heteronomy and a Strong Oral Tradition. Distance Education, 12(2), 163 - 174.

Ebel, R.L. (1972). Essentials of Educational Measurement. New Jersey: Prentice-Hall.

Ebel, R.L., & Frisbie, D.A. (1986). Essentials of Educational Measurement (4th ed.). New Jersey: Prentice-Hall.

Ebel, R.L., & Frisbie, D.A. (1991). Essentials of Educational Measurement (5th ed.). New Jersey: Prentice-Hall.

Englehart, M.D. (1965). A Comparison of Several Item Discrimination Indices. Journal of Educational Measurement, 2, 69 - 76.

Foltz, D. (1990). Toward Better Service and Testing. NHSC Occasional Paper Number 3. Washington, D.C.: National Home Study Council. (ERIC Document Reproduction Service no. ED 323 355).

Frary, R.B. (1991). The None of the Above Option: An Empirical Study. Applied Measurement in Education, 4(2), 115 - 124.

Green, B.G. (1981). A Primer on Testing. American Psychologist, 36, 1001 - 1011.

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

Hambleton, R.K., & Traub, R.E. (1974). The Effect of Item Order on Test Performance and Stress. Journal of Experimental education, 43, 40 - 46.

Harris, B.M., Bessent, W., & McIntyre, K.E. (1969). In Service Education: A Guide to Better Practice. New Jersey: Prentice-Hall.

Henryssen, S. (1971). Gathering, Analyzing, and Using Data on Test Items. In R.L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D.C.: American Council on Education.

Hopkins, K.D., & Stanley, J.C. (1981). Educational and Psychological Measurement and Evaluation (6th ed.). Englewood Cliffs, N.J.: Prentice-Hall.

Huck, S.W., & Bowers, N.D. (1972). Item Difficulty Level and Sequence Effects in Multiple-Choice Achievement. Journal of Educational Measurement, 9(2), 105 - 111.

Kelley, T.L. (1939). The Selection of Upper and Lower Groups for the Validation of test Items. Journal of Educational Psychology, 30, 17 - 24.

Kolstad, R.K., & Kolstad, R.A. (1991). The Option "None of These" Improves Multiple-Choice Test Items. Journal of Dental Education, 55(2), 161 - 163.

Lancaster, D.M. (1987). A Comparison of Item Type and Source on Difficulty and Discrimination Ability. Research Report no. 143. Louisiana, U.S. (ERIC Document Reproduction Service no. ED 290 767).

Lange, A., Lehmann, I.J., & Mehrens, W.A. (1967). Using Item Analysis to Improve Tests. Journal of Educational Measurement, 4(2), 65 - 68.

Lien, A.J. (1976). Measurement and Evaluation of Learning (4th ed.). Dubuque, Iowa: William C. Brown.

Loftus, J. (1988). Writing Examination. In M.P. Lambert and S.R. Welch (Eds.). Home Study Cource Development Handbook (pp. 107 - 128). Washington, D.C.: National Home Study Council. (ERIC Document Reproduction Service no. ED 317 826).

Lord, F.M., & Novick, M.R. (1968). Statistical Theories of Mental Test Scores. Menlo Park, California: Addison-Wesley.

MacKenzie, O., Christensen, E.L., & Rigby, P.H. (1968). Correspondence Instruction in the United States: A of What It Is, How It Functions, and What Its Potential May Be. New York: McGraw-Hill.

Maihoff, N.A., & Mehrens, W.A. (1985). A Comparison of Alternate-Choice and True-False Item forms Used in Classroom Examinations. Report Research no. 143. Michigan, U.S. (ERIC Document Reproduction Service no. ED 269 411).

McMillan, J.R., Mundrake, G.A., & McGuire, S.A. (1989). Multiple-Choice Tests for the Business School: Idealism versus Reality. Delta Pi Epsilon Journal, 31(4), 174 - 181.

Mehrens, W.A., & Lehmann, I.J. (1973). Measurement and Evaluation in Education and Psychology. New York: Holt, Rinehart, & Winston.

Mehrens, W.A., & Lehmann, I.J. (1978). Measurement and Evaluation in Education and Psychology (2nd ed.). New York: Holt, Rinehart, & Winston.

Mehrens, W.A., & Lehmann, I.J. (1984). Measurement and Evaluation in Education and Psychology (3rd ed.). New York: Holt, Rinehart, & Winston.

Monk, J.J., & Stallings, N.M. (1970). Effects of Item Order on Test Scores. Journal of Educational Research, 63, 463 - 465.

Mueller, D.J. (1975). An Assessment of the Effectiveness of Complex Alternatives in Multiple-Choice Achievement test Items. Educational and Psychological Measurement, 35, 135 - 141.

Nunnally, J.C., Jr. (1967). Introduction to Psychological Measurement. New York: McGraw-Hill.

Oosterhof, A.C., & Coats, P.K. (1984). Comparison of Difficulties and Reliabilities of Quantitative Word Problems in Completion and Multiple-Choice Item Formats. Applied Psychological Measurement, 8(3), 287 - 294.

Padilla, M.J., Cronin, L., & Twiest, M. (1985). The Development and Validation of a Test of Basic Process Skills. Paper presented at the annual meeting of the national Association for research in Science teaching, french Lick, Indiana. (ERIC Document Reproduction Service No. ED 256 628).

Payne, D.A. (1974). The Assessment of Learning: Cognitive and Affective. Massachusetts: D.C. Heath and Company.

Plake, B.S. (1980). Item Arrangement and Knowledge of Arrangement on Test Scores. Journal of Experimental Education, 49, 56 - 58.

Plake, B.S., Ansorge, C.J., Parker, C.S., & Lowry, S.R. (1982). Effects of Item Arrangement, Knowledge of Arrangement Test Anxiety and Sex on Test Performance. Journal of Educational Measurement, 19(1), 49 - 57.

Popham, W.J. (1981). Modern Educational Measurement. Englewood Cliffs, N.J.: Prentice-Hall.

Pyrczak, F. (1973). Validity of the Discrimination Index as a Measure of Item Quality. Journal of Educational Measurement, 10(3), 227 - 231.

Sander, A.M. (1988). Scrambled Order-Scrambled Brains: the Effects of Presenting Test Items in Sequential versus Order. Report Research no. 143. Kansas, U.S. (ERIC Document Reproduction Service no. ED 298 169).

Sato, T. (1980). The S-P Chart and the Caution Index. NEC: Educational Informatics Bulletin. Tokyo: Nippon Electric Co., Ltd.

Sax, G. (1974). Principles of Educational Measurement and Evaluation. California: Wadsworth Publishing Company.

Thorndike, R.L. (1951). Reliability. In E.F. Lindquist (Ed.). Educational Measurement (pp. 560-620). Washington, D.C: American Council on Education.

Tinari, F.D. (1979). Item Analysis in Introductory Economics Testing. Improving College and University Teaching, 27(2), 61 - 67.

Tollefson, N. (1987). A Comparison of the Item Difficulty and Item Discrimination of Multiple-Choice Items Using the "None of the Above" and One Correct Response Options. Educational and Psychological Measurement, 47(2), 377 - 383.

Turnbull, W.W. (1956). A Normalized Graphic Method of Item Analysis. Journal of Educational Psychology, 37, 129 - 141.

Wesman, A.G., & Bennett, G.K. (1946). The Use of "None of These" as an Option in Test Construction. Journal of Educational, 37, 541 - 549.

Wood, D.A. (1961). Test Construction: Development and Interpretation of Achievement Tests. Columbus, Ohio: C.E. Merrill.

Zainul, A. (1992). Petunjuk Revisi Butir Soal Universitas Terbuka (The Guideline for Item Revision at the Indonesian Open University, Universitas Terbuka). Unpublished Manuscript, Universitas Terbuka, Exam-Processing Center, Jakarta.

APPENDIX A

Discriminations of Non-revised and Revised Items in FKIP Physisc Program

Item discrimination in each course											
PFIS4439				PFIS4438				PFIS4437			
Non-revised		Revised		Non-revised		Revised		Non-revised		Revised	
First use	Second use	Before revised	After revised	First use	Second use	Before revised	After revised	First use	Second use	Before revised	After revised
0.327	0.392	0.221	0.368	0.399	0.412	0.186	-0.074	0.246	0.100	0.257	0.317
0.398	0.199	0.104	0.397	0.452	0.293	0.165	-0.033	0.576	0.504	0.368	0.246
0.455	0.358	0.318	0.451	0.279	0.091	0.086	-0.051	0.386	0.437	0.015	0.135
0.514	0.316	0.494	0.431	0.241	0.244	0.122	0.156	0.354	0.394	0.397	0.376
0.387	-0.076	0.473	0.454	-0.094	0.128	-0.005	0.140	0.395	0.460	0.191	0.239
0.338	0.471	0.237	0.385	0.071	0.119	0.300	0.147	0.558	0.275	-0.013	0.244
0.257	0.064	0.083	0.294	0.266	0.251	0.112	0.126	0.366	0.299	-0.189	0.164
0.393	0.228	0.331	0.123	0.348	0.240	0.252	0.321	0.329	0.422	0.130	0.323
0.352	0.483	0.165	0.318	0.155	0.058	0.182	0.049	0.274	0.358	0.175	0.305
0.313	0.414	0.158	-0.120	0.385	0.237	0.061	0.209	0.397	0.339	-0.029	0.381
0.327	0.208	0.028	0.157	0.212	0.295	0.080	0.079	0.304	-0.014	0.180	0.061
0.176	0.213	0.143	0.191	0.319	0.256	0.412	0.365	0.219	0.228	0.246	0.114
0.223	0.264	0.342	0.365	0.316	0.295	0.145	0.265	0.239	0.335	0.169	0.308
0.185	0.296	0.347	-0.139	0.406	0.316	0.176	0.114	0.500	0.244	0.161	0.205
0.427	0.399	0.342	0.397	0.345	0.286	0.208	0.191	-0.004	0.116	0.317	0.374
0.230	0.311	0.398	0.215	0.345	0.267	0.190	0.305	0.212	0.176	0.436	0.257
0.302	0.267	0.232	0.351	0.332	0.302	0.110	0.114	0.303	-0.053	0.202	0.234
0.324	0.307	0.180	0.125	0.469	0.323	0.289	0.002	0.198	0.362	0.147	0.100
0.117	0.345	0.025	0.348	0.383	0.170	0.069	0.067	0.408	0.276	0.050	0.225
0.358	0.247	0.104	0.289	0.223	0.191	0.011	0.144	0.465	0.383	0.064	0.036
0.159	0.313	0.108	0.269	0.241	0.367	0.190	0.279	0.148	0.301	0.068	0.057
0.309	0.291	0.302	0.358	0.206	0.279	0.162	0.272	0.353	0.373	0.210	0.153
0.200	-0.017	0.168	0.100			-0.099	-0.151	0.216	0.050		
		0.051	0.147			0.129	0.149	0.481	0.223		
		0.246	-0.047			0.203	0.228	0.340	0.139		
		0.173	0.255			0.170	0.128	0.288	0.223		
						0.195	0.074	0.523	0.312		
						0.156	0.151				

APPENDIX B

Discriminations of Non-revised and Revised Items in FKIP English Program

Item discrimination in each course											
PING4447				PING4441				PING4448			
Non-revised		Revised		Non-revised		Revised		Non-revised		Revised	
First use	Second use	Before revised	After revised	First use	Second use	Before revised	After revised	First use	Second use	Before revised	After revised
-0.099	-0.265	0.233	0.370	0.302	0.207	0.153	0.350	0.244	0.241	0.297	0.234
0.242	0.227	0.272	0.345	0.283	0.263	0.047	0.069	-0.029	-0.013	0.212	0.240
0.236	0.008	0.113	0.103	0.388	0.313	0.311	0.221	0.323	-0.017	0.311	0.144
0.420	0.152	0.060	0.144	0.361	0.392	0.282	0.427	0.247	0.111	0.357	0.197
0.498	0.441	0.280	0.332	0.284	0.492	0.275	0.147	0.555	0.349	0.129	0.188
0.221	0.361	0.188	0.303	0.271	0.247	0.164	0.451	0.367	0.381	0.410	0.397
0.408	0.250	0.396	0.339	0.262	0.240	0.071	0.076	0.425	0.241	0.354	0.289
0.551	0.370	0.087	0.376	0.191	0.402	0.553	0.449	0.345	0.114	0.037	0.195
0.494	0.353	0.345	0.369	0.529	0.502	0.168	0.352	0.296	0.210	0.379	0.178
0.364	0.192	0.284	0.160	0.203	0.153	0.230	0.228	0.412	0.345	0.253	0.204
0.285	0.174	0.325	0.160	0.213	0.396	0.504	0.465	0.512	0.430	0.137	0.098
0.364	0.256	0.279	0.255	0.217	0.426	0.256	0.191	-0.006	-0.079	0.274	0.254
0.032	0.144	0.207	0.214	0.558	0.492	-0.082	0.000	0.170	0.178	0.183	0.230
0.386	0.518	0.050	0.062	0.464	0.398	0.258	0.518	0.307	0.405	0.503	0.411
0.546	0.500	0.220	0.374	0.243	0.325	0.008	0.194	-0.085	0.149	0.459	0.432
0.381	0.397	0.100	0.371	0.404	0.331	0.045	0.129	0.418	0.000	0.473	0.489
0.398	0.369	0.129	0.280	0.039	0.240	0.066	0.256	0.245	0.293	0.452	0.148
0.419	0.110	0.141	0.483	0.401	0.521	0.100	0.288	0.525	0.626	0.512	0.440
0.483	0.436	-0.286	0.265	0.296	0.361	0.128	0.144	0.290	0.156	0.048	-0.052
0.478	0.259	0.186	0.091	0.354	0.346	0.175	0.163	0.227	0.224	0.316	0.320
0.219	0.173	0.104	0.120	0.349	0.338	0.235	0.452	0.350	0.207	0.477	0.335
0.332	0.345	0.226	-0.006	0.376	0.499	-0.060	0.564	0.446	0.259	0.210	0.172
0.470	0.390	0.105	-0.027	0.446	0.392	0.337	0.431	0.444	0.395	0.356	0.364
0.280	-0.103	0.106	0.129	0.269	0.093	0.058	0.371	0.367	0.407	0.391	0.279
0.419	0.226	0.194	0.135	0.281	0.261	0.068	0.256	0.454	0.332	0.276	0.358

table continues...

Item discrimination in each course

PING4447				PING4441				PING4448			
Non-revised		Revised		Non-revised		Revised		Non-revised		Revised	
First use	Second use	Before revised	After revised	First use	Second use	Before revised	After revised	First use	Second use	Before revised	After revised
0.430	0.156	0.299	0.251	0.112	0.188	0.237	-0.099	0.448	0.376		
0.418	0.325	0.139	0.168	0.426	0.426	0.128	0.242	0.307	0.190		
0.296	0.326	0.108	0.226	0.502	0.417	0.097	0.268	0.398	0.489		
0.457	0.381	0.166	0.119	0.361	0.260	-0.013	0.546	0.302	0.342		
0.232	0.385			0.488	0.449	0.443	0.280	0.220	0.240		
0.309	0.263							0.497	0.460		
0.319	0.259							0.033	0.287		
0.410	0.229							0.407	0.392		
0.299	0.349							0.556	0.513		
0.361	0.490							0.501	0.536		
0.453	0.437										
0.210	0.066										
0.431	0.101										
0.305	0.369										
0.471	0.474										
0.145	0.389										

APPENDIX C

Discriminations of Non-revised and Revised Items in FMIPA Mathematics Program

Item discrimination in each course							
MATK4110				MATK4112			
Non-revised		Revised		Non-revised		Revised	
First use	Second use	Before revised	After revised	First use	Second use	Before revised	After revised
0.182	0.465	0.219	0.577	0.223	0.174	0.022	0.035
0.448	0.521	0.194	0.120	0.261	0.388	0.326	0.158
0.171	0.078	0.182	0.494	0.262	0.347		
0.488	0.548	0.448	0.575	0.365	0.420		
0.385	0.394	0.174	-0.046	0.401	0.448		
0.325	0.364	0.115	0.215	0.096	0.117		
0.365	0.381	0.156	0.205	0.134	0.183		
0.216	0.249	0.211	0.244	0.118	0.050		
0.331	0.428	0.043	0.095	0.590	0.246		
0.292	0.103	0.055	-0.002	0.174	0.243		
0.000	0.059	0.475	0.562	0.161	0.230		
0.176	0.191	0.131	0.108	0.472	0.394		
0.085	0.108	0.331	0.325	0.519	0.479		
0.288	0.499			0.353	0.445		
0.191	0.249			0.403	0.360		
0.382	0.279			0.204	0.177		
0.367	0.421			0.139	0.114		
				0.340	0.265		
				0.441	0.315		
				0.388	-0.088		
				0.375	0.322		
				0.209	0.174		
				0.643	0.271		
				0.537	0.076		
				0.370	-0.076		
				0.463	0.322		
				0.199	0.315		

APPENDIX D

Discriminations of Non-revised and Revised Items in FMIPA Statistics Program

Item discrimination in each course											
STAT4110				STAT4213				STAT4334			
Non-revised		Revised		Non-revised		Revised		Non-revised		Revised	
First use	Second use	Before revised	After revised	First use	Second use	Before revised	After revised	First use	Second use	Before revised	After revised
0.247	0.319	0.228	0.158	0.176	0.000	0.173	-0.003	0.205	0.550	0.098	0.490
0.253	0.332	0.190	0.141	0.278	0.545	0.278	0.406	0.308	0.157	0.270	0.564
0.138	0.303	0.345	0.257	0.395	0.163	0.134	0.483	0.060	0.276	0.163	0.083
0.389	0.201	0.194	0.257	0.235	0.215	0.042	-0.022	0.292	0.626	0.177	-0.112
0.323	0.234	0.222	0.105	0.445	0.382	0.211	0.077	0.289	0.674	0.237	0.014
0.294	0.224	0.114	0.078	0.242	0.280	0.086	-0.028	0.482	0.412		
0.266	0.257	0.187	0.181	0.374	0.015			0.200	-0.157		
0.128	0.303	0.263	0.214	0.043	0.215			0.209	0.171		
0.316	0.398			0.304	0.391			0.335	0.307		
0.263	0.263			0.381	0.360			0.580	0.512		
0.388	0.072			-0.032	0.234			0.274	0.398		
0.360	0.224			0.089	0.080			0.098	0.550		
0.256	0.214			0.206	0.203			0.125	0.507		
0.203	0.401			0.071	0.172			0.161	0.102		
0.266	0.227			0.275	0.215			0.223	0.417		
0.342	0.224			0.348	0.237			0.256	0.131		
0.184	0.270			0.332	0.538			0.223	0.398		
0.212	0.155			0.003	0.542			0.260	0.240		
0.155	0.092			0.302	-0.015			0.417	0.521		
0.311	0.168			0.310	0.369			0.387	0.607		
0.062	0.299			0.297	0.268			0.172	0.081		
0.301	-0.036			0.150	0.255			0.181	0.205		
				0.246	0.154			0.177	-0.162		
				0.284	-0.012			0.510	0.119		
								0.512	0.007		

APPENDIX E

Illustration of Item Revision

Statistical data of students' responses before and after revision for eight items in Calculus I are indicated in the following paragraphs.

The following four items have been written in the form of usual multiple-choice (type A) questions which ask students to select one correct answer from four options.

Item 1.

Before Revision (in test 89.1).

If $f(x) = 3^x$ then $f(x + 3) - f(x - 1)$ equal to...

A. 3^x (27 - 1/9). (p = 0.357, D = - 0.076)

B. 3^x (9 - 1/3). (p = 0.381, D = - 0.109)

* C. 80 (3^{x-1}). (p = 0.167, D = + 0.219)

D. 80 (3^{x+1}). (p = 0.071, D = - 0.106)

OMMIT (p = 0.024, D = + 0.201)

After Revision (in test 92.1).

If $f(x) = 3^x$ then $f(x + 3) - f(x - 1)$ equal to...

A. 3^x (27 - 1/9). (p = 0.065, D = - 0.046)

B. 3^x (9 - 1/3). (p = 0.286, D = - 0.430)

* C. 3^x (27 - 1/3). (p = 0.494, D = + 0.577)

D. 3^x (9). (p = 0.156, D = - 0.218)

OMMIT (p = 0.000, D = 0.000)

Before revision, the item was somewhat too difficult for the group tested (only 16.7% of students choose the correct answer) and did not discriminate well (the item was categorized as marginal). The distractors functioned well. However, 2.4% of students did not answer the item. After revision, the item was much easier (49.4% of students chose the correct answer) and much more highly discriminating than the original item (the item was categorized as very good). The distractors also functioned better than those in the original items with no students ommitting the item.

Item 2.

Before Revision (in test 88.2).

Determine the area of the rectangle with its base on x-axis and its upper corners on curve $y = 12 - x^2$.

A. 30 (p = 0.202, D = - 0.118)

* B. 32 (p = 0.495, D = + 0.194)

C. 35 (p = 0.083, D = + 0.076)

D. 37 (p = 0.202, D = - 0.118)

OMMIT (p = 0.018, D = - 0.188)

<u>After Revision</u> (in test 92.1).

Determine the maximum area of the rectangle with its base on x-axis and its upper corners on curve $y = 12 - x^2$.

A. 30 (p = 0.234, D = - 0.169)

* B. 32 (p = 0.494, D = + 0.120)

C. 35 (p = 0.143, D = + 0.005)

D. 37 (p = 0.117, D = + 0.010)

OMMIT (p = 0.013, D = + 0.046)

Both before and after revision, the item was categorized as moderate with a poor discrimination. Both before and after revision, distractor C did not function well. Distractor D functioned well in the original item (before revision), but it did not function well in the revised items (after revision).

Item 3.

Before Revision (in test 89.1)

Given $y = \text{arc cotg } \frac{1 + x}{1 - x}$. Then $dy/dx = \ldots$

A. $\frac{2x}{1 + x^2}$ (p = 0.095, D = - 0.219)

B. $\frac{2x}{1 - x^2}$ (p = 0.214, D = + 0.049)

C. $\frac{1}{1 - x^2}$ (p = 0.321, D = - 0.091)

* D. $-\frac{1}{1 + x^2}$ (p = 0.369, D = + 0.182)

OMMIT (p = 0.000, D = 0.000)

After Revision (in test 92.1)

$y = \text{arc cotg } \frac{1 + x}{1 - x}$, $dy/dx = \ldots$

A. $-\frac{1}{1 - x^2}$ (p = 0.156, D = - 0.262)

* B. $-\frac{1}{1 + x^2}$ (p = 0.636, D = + 0.494)

C. $-\frac{1 + x^2}{1 - x^2}$ (p = 0.104, D = - 0.352)

D. $-\frac{2x}{1 + x^2}$ (p = 0.091, D = - 0.066)

OMMIT (p = 0.013, D = - 0.149)

Before revision, the item was somewhat difficult for the group tested (36.9% of students selected the correct answer) and had a poor discrimination. One option, option B, did not function well. After being revised, the item was much easier and much more highly discriminating than the original item (the item was categorized as moderate with very good discrimination). The distractors also functioned better than those in the original item.

Item 4.

Before Revision (in test 88.2).

The length of curve $\rho = \theta^2$ from $\theta = 0$ to $\theta = 2\sqrt{3}$ is.....

	A. 55/2	(p = 0.174, D = - 0.091)
	B. 55/4	(p = 0.193, D = + 0.138)
*	C. 56/3	(p = 0.505, D = + 0.156)
	D. 56/5	(p = 0.083, D = - 0.168)
	OMMIT	(p = 0.046, D = - 0.244)

After Revision (in test 92.1)

The length of curve $\rho = \theta^2$ from $\theta = 0$ to $\theta = 2\sqrt{3}$ is ......
(remember the formula for the length of curve in a polar coordinate)

	A. 55/2	(p = 0.234, D = - 0.122)
	B. 55/4	(p = 0.156, D = + 0.046)
*	C. 56/3	(p = 0.481, D = + 0.205)
	D. 56/5	(p = 0.104, D = - 0.149)
	OMMIT	(p = 0.026, D = - 0.171)

Before revision, the item was categorized as moderate with poor discrimination.

After revision, the item was still categorized as moderate but the discrimination shifted to the marginal category.

Both before and after revision, distractor B did not function well.

The following four items have been written in the form of multiple multiple-choice (type D) questions which ask students to select: A if options 1) and 2) are correct; B if options 1) and 3) are correct; C if options 2) and 3) are correct; and D if options 1), 2), and 3) are correct.

Item 5.

<u>Before Revision</u> (in test 89.1).

A function $f(x) = \begin{cases} 4 + x, & \text{if} \quad x < 2 \\ 4 - x, & \text{if} \quad x \geqslant 2 \end{cases}$

1) the function is discontinuous at $x = 2$.

2) $\lim_{x \longrightarrow 2^+} f(x) = 2.$

3) $\lim_{x \longrightarrow 2} f(x)$ exist.

Students' responses:

	Difficulty (p)	Discrimination (D)
* A.	0.286	+ 0.055
B.	0.071	- 0.009
C.	0.286	+ 0.188
D.	0.357	- 0.225
OMMIT	0.000	0.000

<u>After Revision</u> (in test 92.1).

A function $f(x) = \begin{cases} 4 + x, & \text{if } x < 2 \\ 4 - x, & \text{if } x \geq 2 \end{cases}$

A statement that is fulfilled by f(x) is .....

1) the function is discontinuous at x = 2.

2) $\lim_{x \to 2^+} f(x) = 2.$

3) $\lim_{x \to 2} f(x)$ exist.

Students' responses:

	Difficulty (p)	Discrimination (D)
* A.	0.260	- 0.002
B.	0.156	- 0.098
C.	0.273	+ 0.235
D.	0.273	- 0.098
OMMIT	0.039	- 0.120

Both before and after revision, the item was somewhat too difficult for the group tested and did not discriminate well. Both before and after revision, some good students were more interested in distractor D than in the correct answer.

Item 6.

Before Revision (in test 88.1)

Function f(x) is called continuous at point $x = x_0$ if

1) $f(x_0)$ is defined.

2) $\lim_{x \to x_0} f(x)$ exist.

3) $\lim_{x \to x_0} f(x) = f(x_0)$.

Students' responses:

	Difficulty (p)	Discrimination (D)
A.	0.196	- 0.235
B.	0.141	- 0.239
C.	0.067	- 0.209
* D.	0.595	+ 0.475
OMMIT	0.000	0.000

After Revision (in test 92.1).

If function f(x) is continuous at point $x = x_0$, then

1) $\lim_{x \to x_0^-} f(x)$ exist.

2) $\lim_{x \to x_0^+} f(x)$ exist.

3) $\lim_{x \to x_0^-} f(x) = \lim_{x \to x_0^+} f(x) = f(x_0)$.

Students' responses:

	Difficulty (p)	Discrimination (D)
A.	0.078	- 0.281
B.	0.078	- 0.191
C.	0.169	- 0.293
* D.	0.636	+ 0.562
OMMIT	0.039	- 0.152

Both before and after revision, the item was categorized as moderate with a very high discrimination. Both before and after revision, the distractors functioned well.

Item 7.

Before Revision (in test 89.1).

Function $f(x) = x^3 + 3x$

1) increases for all values of x.
2) increases in interval $x < -3$.
3) does not have extreme values.

Students' responses:

	Difficulty (p)	Discrimination (D)
A.	0.131	- 0.167

*	B.	0.548	+ 0.131
	C.	0.202	- 0.198
	D.	0.107	+ 0.249
OMMIT		0.012	- 0.055

<u>After Revision</u> (in test 92.1).

Function $f(x) = x^3 + 3x$, then

1) f(x) increases for all values of x.
2) f(x) increases in interval $x < -3$.
3) f(x) does not have extreme values.

Students' responses:

		Difficulty (p)	Discrimination (D)
	A.	0.195	- 0.078
*	B.	0.481	+ 0.054
	C.	0.143	- 0.076
	D.	0.156	+ 0.108
OMMIT		0.026	- 0.068

Both before and after revision, the item was categorized as moderate with a poor discrimination. Both before and after revision, some good students were more interested in distractor D than in the correct answer.

Item 8.

Before Revision (in test 89.1).

A particle moves on a line $s = t^3 - 6t^2 + 9t + 4$.

Determine S and v if a = 0.

1) S = 6.

2) v = -3.

3) t = 2.

Students' responses:

		Difficulty (p)	Discrimination (D)
	A.	0.333	+ 0.006
	B.	0.262	- 0.246
	C.	0.071	- 0.191
*	D.	0.333	+ 0.331
OMMIT		0.000	0.000

After Revision (in test 92.1).

A particle moves on a line $s = t^3 - 6t^2 + 9t + 4$.

If a = 0 then

1) S = 6.

2) v = -3.

3) t = 2.

Students' responses:

	Difficulty (p)	Discrimination (D)
A.	0.143	- 0.176
B.	0.364	- 0.115
C.	0.182	- 0.105
* D.	0.299	+ 0.325
OMMIT	0.013	+ 0.071

Both before and after revision, the item was categorized as difficult with a reasonably good discrimination. Before revision, distractor A did not function well but after revision, it was.

APPENDIX F

Table of Specifications of Final Examinations at the Indonesian Open University

Faculty : Mathematics and Natural Science
Program : Mathematics
Course : Calculus I
Writers : N. Soemartojo and Djati K.

No.	Topics and objectives	Cognitive levels C1,2	C3	C4,5	C6	Total
1.	THE SET OF NUMBERS					
	A. Line diagram of a set of numbers	-	-	-	-	-
	B. Line of real numbers	1	-	-	-	1
	C. Interval	-	1	-	-	1
2.	PERMUTATIONS					
	A. Complete Inductions	-	-	-	-	-
	B. Permutations and Combinations	-	-	2	-	2
	C. Binomium Newton	1	-	-	-	1
3.	FUNCTIONS					
	A. Understanding of Functions	-	-	1	-	1
	B. Limits and Continuities	-	-	-	1	1
	C. The Derivatives	-	1	-	-	1
4.	MAXIMA AND MINIMA					
	A. Algebraic-function differentiation	1	-	-	-	1
	B. Exponential-function differentiation	-	-	-	1	1
	C. Critical points	-	-	-	1	1

table continues...

No.	Topics and objectives	Cognitive levels C1,2	C3	C4,5	C6	Total
5.	FUNCTION DIFFERENTIATION					
	A. Trigonometric-function differentiation	-	1	-	-	1
	B. Exponential-function differentiation	-	2	-	-	2
	C. Applications of differentiation	-	-	1	-	1
6.	PARAMETRIC FUNCTIONS					
	A. Polar coordinates	-	-	1	-	1
	B. Curvature	-	1	-	-	1
	C. Evolutes	-	1	-	-	1
7.	THE MEAN VALUE THEOREM					
	A. Rolle's theorem	-	-	1	-	1
	B. The mean value theorem	-	1	-	-	1
	C. Indeterminate forms	1	-	-	-	1
8.	INTEGRALS					
	A. Algebraic and trigonometric-function integrals	1	-	1	-	2
	B. Partial integrals and reduction formula	1	-	-	-	1
	C. Rational-function integrals	-	1	-	-	1
9.	AREA AND DEFINITE INTEGRAL					
	A. Definite integral	-	-	1	-	1
	B. The problem of area	-	1	-	1	2
	C. The problem of area in polar coordinate	-	1	1	-	2

table continues...

No.	Topics and objectives	Cognitive levels C1,2	C3	C4,5	C6	Total
10.	VOLUMES OF SOLIDS OF REVOLUTION					
	A. Disc method	-	1	-	-	1
	B. Shell method	-	2	-	-	2
	C. The first theorem of Pappus	-	1	1	-	2
11.	AREA SURFACES OF REVOLUTION					
	A. Arc length	1	-	1	-	2
	B. Area of surfaces of revolution	-	1	1	-	2
	C. The second theorem of Pappus	1	-	-	-	1
12.	APPLICATION OF CALCULUS					
	A. Calculus and Physics	-	1	1	-	2
	B. Calculus and Technology	-	-	-	-	-
	C. Calculus and Industry	-	-	2	1	3
	T o t a l	8	17	15	5	45
	P e r c e n t a g e	20%	40%	30%	10%	100%

APPENDIX G

Table of Specifications of Final Examinations at the Indonesian Open University

Faculty : Mathematics and Natural Science
Program : Statistics
Course : Statistical Method I

Writer: Zanzawi Soejoeti

No	Topics and objectives	Cognitive levels C1	C2	C3	C4,5,6	Total
I.	1. Able to explain the meaning of statistics	1	-	-	-	1
	2. Able to describe and explain the scales of measurement	-	1	-	-	1
	3. Able to use the "sigma" notation	-	-	1	-	1
II.	1. Able to present data					
	2. Skilled in calculating the measures of central tendency	-	-	2	-	2
	3. Skilled in calculating the measures of variability	-	-	1	-	1
III.	1. Able to explain the sample space and event					
	2. Able to explain and calculate the probability of an event	-	-	2	-	2
	3. Able to explain and calculate the conditional probability	-	-	1	-	1

table continues...

No	Topics and objectives	Cognitive levels C1	C2	C3	C4,5,6	Total
IV.	1. a. Recognize the result of quantitative experiment as a random variable b. Recognize the probability distribution	-	-	2	-	2
	2. Skilled in calculating the expected value, variance, and in describing their characteristics					
	3. a. Able to explain and calculate the joint distributions of two random variables b. Able to explain the relationship between two random variables and to calculate the the covariance and correlation coefficient	-	-	2	-	2
V.	1. Recognize and be able to use Bernoulli model	-	-	2	-	2
	2. Recognize and be able to use Binomial distribution					
	3. Recognize and be able to use Hypergeometric and Poisson distributions	-	-	2	-	2
VI.	1. Recognize and be able to calculate the probability based on the Normal distribution	-	-	2	1	3
	2. Be able to explain the concept of sampling distribution and to recognize their characteristics	-	-	1	-	1

table continues...

No	Topics and objectives	Cognitive levels C1	C2	C3	C4,5,6	Total
	3. Be able to apply the Normal distribution approach to Binomial distribution	-	-	1	-	1
VII.	1. Be able to explain and calculate the point and interval estimations	-	1	1	-	2
	2. a. Be able to formulate the null hypothesis for a problem					
	b. Be able to explain and use the basic concept of type II error	-	-	-	-	-
	c. Be able to calculate the significance level and power of test					
VIII.	1. Be able to estimate and test the hypotheses for proportion and mean of population using a large sample	-	-	1	1	2
	2. a. Be able to use chi-square and t distributions					
	b. Be able to estimate and test the hypotheses for mean parameter and variance of a Normal population	-	-	2	1	3
IX.	1. Be able to estimate and test the hypotheses for difference of proportions and means from two arbitrary populations using a large sample	-	-	1	1	2

table continues...

No	Topics and objectives	Cognitive levels C1	C2	C3	C4,5,6	Total
	2. a. Be able to use F distribution b. Be able to test hypotheses for equality of variances and means of two Normal populations c. Be able to estimate the ratio of variances of two Normal populations d. Be able to draw an inference for paired dependent samples from a Normal population	-	-	2	1	3
	T o t a l	1	2	26	5	34
	P e r c e n t a g e	3%	6%	76%	15%	100%

APPENDIX H

Example of Item Analysis

KODE MATA KULIAH MATK4110

TGL.UJIAN : 17 APRIL 88
KODE NASKAH : 014
JUMLAH SAMPLE MHS : 163

MT : 12.40
SD : 004.69
KR-20 : +0.719
SEM : 02.485

SOAL	PROPORSI YANG MENJAWAB C	A	B	C	D	E	KUNCI	P	Q	MP	R-BIS
1	0.006	0.153	0.209	0.423	0.209	0.000	D	0.209	0.791	14.44	+0.222
R-BIS	-0.136	-0.128	-0.190	+0.087	+0.222	+0.000					
MEAN	06.00	10.96	10.65	12.88	14.44	00.00					
2	0.006	0.166	0.466	0.117	0.245	0.000	B	0.466	0.534	14.26	+0.369
R-BIS	-0.136	-0.077	+0.369	-0.156	-0.224	+0.000					
MEAN	06.00	11.59	14.26	10.37	10.55	00.00					
3	0.006	0.313	0.092	0.423	0.166	0.000	C	0.423	0.577	13.29	+0.162
R-BIS	+0.055	+0.002	-0.124	+0.162	-0.130	+0.000					
MEAN	15.00	12.41	10.60	13.29	11.04	00.00					
4	0.000	0.086	0.761	0.049	0.104	0.000	B	0.761	0.239	13.31	+0.345
R-BIS	+0.000	-0.134	+0.345	-0.235	-0.183	+0.000					
MEAN	00.00	10.29	13.31	07.38	09.94	00.00					
5	0.000	0.110	0.595	0.252	0.043	0.000	B	0.595	0.405	14.43	+0.525
R-BIS	+0.000	-0.237	+0.525	-0.403	-0.036	+0.000					
MEAN	00.00	09.22	14.43	09.15	11.57	00.00					
6	0.006	0.141	0.104	0.650	0.098	0.000	C	0.650	0.350	13.54	+0.330
R-BIS	-0.072	-0.156	-0.145	+0.330	-0.181	+0.000					
MEAN	09.00	10.57	10.47	13.54	09.81	00.00					
7	0.000	0.282	0.436	0.055	0.227	0.000	D	0.227	0.773	15.76	+0.386
R-BIS	+0.000	-0.119	-0.179	-0.077	+0.386	+0.000					
MEAN	00.00	11.50	11.44	10.89	15.76	00.00					

VITA

Surname: Pakpahan Given Names: Sondang Purnamasari

Place of Birth: Jakarta, Indonesia Date of Birth: September 11, 1962

Educational Institutions Attended:

University of Indonesia, Jakarta.	1981 to 1987
University of Victoria, B.C.	1990 to 1993

Degrees Awarded:

B.Sc.	University of Indonesia	1987

Honours and Awards:

Publications:

_____

_____

_____

_____

_____

PARTIAL COPYRIGHT LICENSE

I hereby grant the right to lend my thesis to users of the University of Victoria Library, and to make single copies only for such users or in response to a request from the Library of any other University, or similar Institution, on its behalf or for one of its users. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by me or a member of the University designated by me. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Title of Thesis: Comparison of Test Items Within and Between Faculties at the Indonesian Open University

Author 

SONDANG PURNAMASARI PAKPAHAN

April 8, 1993