AN ORDERING THEORETIC ANALYSIS OF THE SATO CAUTION INDICES IN A MALAYSIAN CONTEXT

By

Ivan Douglas Filmer, Jr.

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational
Psychology and Special Education

1992

ABSTRACT

AN ORDERING THEORETIC ANALYSIS OF THE SATO CAUTION INDICES IN A MALAYSIAN CONTEXT

By

Ivan Douglas Filmer, Jr.

The Sato caution indices are derived by investigating the observable patterns of students' responses on a test and summary statistics. It involves constructing a students-by-items matrix of the binary responses where students are arranged from highest to lowest scoring, and items arranged in order of increasing difficulty. The indices derived by a formula introduced by Sato ranges from zero to over 1 and indicate the extent to which items are aberrant.

In this study, a 40-item objective test was administered to 354 fifth form students in six Malaysian schools. The item and student caution indices were calculated first using the students-by-items matrix of the Sato model and then again after ordering the items with the probabilistic model Z of the ordering theory (Krus, 1975). A principal components factor analysis and an agglomerative hierarchical cluster analysis were conducted on the items to aid the ordering theoretic analysis.

The results of the study showed that in the Sato model, the ordering or arrangement of the items did not affect the

calculation of the item caution indices but affected the calculation of the student caution indices. Similarly, the ordering of the students affected the calculation of the item caution indices. The sample size of items and students had an effect on the magnitudes of the item and student caution indices derived.

The arrangement of the items according to the ordering theoretic analysis correlated almost identically with the arrangement of items in the Sato model (r_{ranks}=.999). Identical item caution indices were produced. The item caution indices derived from the different group characteristics of item format, school location, students' gender, students' SES, and teachers' working experience, were all not significantly different. There was also no significant interaction effect between the student caution indices derived from students of different SES and different school locations.

Dedicated to my wife,

Voon Mooi,

and my children,

Andrew and Andrea.

ACKNOWLEDGEMENTS

This dissertation has been completed with the help of a great many people. First of all, I wish to thank the Malaysian Ministry of Education for awarding me a scholarship to aid me in my doctoral studies, the Assistant Director of the Malaysian Educational Planning and Research Division, Dr. Hanafi Mohamed Kamal, and Tuan Haji Jumali Kassan, the State Education Director of Selangor Darul Ehsan.

Special thanks are extended to Puan Nik Faizah Nik Mustapha, the former Assistant Director of the Examinations Syndicate, for her genuine concerns, Mr. Lim Chee Tong, the former Assistant Director of the Vocational and Technical Schools Division, without whose help this study would have been more difficult, and Puan Hajah Badiah Abdul Manan, the former Director of the Examinations Syndicate, for granting permission to obtain certain data from the Syndicate.

I am also in debt to Datin Hajah Rapiah Tun Abdul Aziz, En. Abdul Rafor Ibrahim and Mr. A. Sivanesan, all senior officers of the Examinations Syndicate. I wish to express my appreciation also to my friend, Mr. Leslie Fredericks, the Head of the Examinations Unit at the Teachers' Training Division, for helping me at some of the crucial stages of my study. Without his help, this study would have taken a much

longer time to complete.

I wish to express my gratitude to the members of my Dissertation Committee. I wish to thank Dr. William Mehrens, who served as the dissertation chairman of my committee, for his concerns and guidance in the completion of this study. His friendship and willingness to offer suggestions are most appreciated. I also wish to thank Dr. Betsy Becker for her statistical expertise and suggestions in the data analysis, Dr. Norman Bell for his constructive suggestions as a member of the committee, and Dr. Frederick Ignatovich for his assistance during the initial stages of my study.

I especially appreciate the support and cooperation of all the pupils, teachers and school administrators who participated in this study. I will always remember their contributions of time and effort.

Finally and most of all, I wish to thank my wife, Voon Mooi, for her love, patience, understanding and support. I appreciate the many sacrifices made by her and my children, Andrew and Andrea, to make it all possible.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

STATEMENT OF THE PROBLEM

Introduction

It has been common practice in education to use a single total score on a test to report the academic achievement or ability of a student. However, Harnisch and Linn (1981) have pointed out that there are as many as 184,756 possible item response patterns that yield a score of 10 on a 20-item test. Therefore, a student's total test score, which is merely the number of items correct on a test, can often be misleading. Blixt and Dinero (1985) echo the view of many educational authorities when they pointed out, "There is no guarantee that any one total score on a test will give the same information about each of the examinees who take the test" (p. 239). As such, in order to assess student performance and student errors, teachers or administrators need more than a single total test score (Harnisch, 1983). Presently, researchers have turned to examining more closely the item response patterns of students. The type of information obtained from such analyses is above and beyond the kind provided by the traditional method of scoring a test, which is, as Birenbaum and Shaw (1985) explain, the consideration only for the total number of correct responses.

Researchers like Birenbaum & Tatsuoka (1982), and Tatsuoka & Tatsuoka (1983) have shown that the analysis of students' response patterns in a test would enable a teacher to prescribe individualized remediation to correct some of those students' misconceptions. Others like Brown & Burton (1978), and Tatsuoka & Baillie (1982) have developed computer programs, such as "BUGGY" and "SIGNBUG", for diagnosing students' misconceptions of learning from tests.

However, the appropriate interpretation of any test depends on knowing the dimensions underlying the items and the correspondence between the items and those dimensions (Green, 1983). Studies have also shown that tests that were previously calibrated as unidimensional were subsequently found to be multidimensional (Reckase, 1985; Bock, Gibbons, & Muraki, 1986; Zimowski & Bock, 1987). If a test was multidimensional, it would be even more inaccurate to say that two students receiving the same score on a test possess the same ability or have attained the same level of academic achievement. This problem has stimulated research in different approaches to diagnosing atypical item response patterns of tests.

Studies have shown that a variety of factors have contributed to atypical item response patterns. Prominent among these factors are differences in students' background experiences, students' exposure to different subject matter or school-to-school variability in content coverage and emphases, test anxiety, students' guessing, carelessness, cheating on the test, different curriculum emphases or coverage, and

students' attendance patterns.

As many as 20 different types of item response indices have been formulated to gauge the extent to which an individual's response pattern on a test is unusual. Harnisch and Linn (1981) have conveniently categorized these indices into two major groups. The first group, which they label "appropriateness" indices, is based on the item response theory (IRT). The second is based directly on the observable pattern of right and wrong answers, and summary statistics. The most popular index used in this latter group is the Sato caution index.

Problem Statement

The problem addressed in this study concerns the calculation of the Sato caution indices. These indices are derived based on the assumption that the items on a test may be linearly arranged in terms of item difficulty. A students-by-items matrix is first constructed wherein students are arranged from highest scoring to lowest, and items are arranged from easiest to most difficult. As most tests are generally found to be multidimensional, this simple ordering may not be adequate. As such, the values of the caution indices derived in this way may not be accurate. Furthermore, when arranging the matrix in this manner, the ordering of marginal totals that have the same magnitude, are resolved arbitrarily. This is to say that there is no organized or conceptual method of arranging the students who have the same

score, or items that have the same p-value. According to McArthur (1987), these arbitrary allocations contribute to the instability of the S (Student) and P (Problem) curves in the students-by-items matrix. Invariably, this instability would affect the accuracy of the caution indices.

Harnisch (1983) has stated that a student caution index is group dependent because it indicates whether an individual student's pattern of responses is atypical, relative to the responses of the whole group of students who took the test. No research work has been reported on the effects of the group characteristics on the caution indices.

The Purpose of the Study

The purpose of this study is to address two main issues concerning the Sato caution index. The first issue is the effect of the dimensionality of the test on the calculation of the caution indices. The second issue is the type of group characteristics affecting the magnitudes of the item and student caution indices.

Concerning the first issue, the purpose is to apply the ordering theoretic analysis of Bart and Krus (1973) on the students' responses to a test instrument to determine the hierarchical pattern of items on the test. The test items will then be arranged according to the hierarchical pattern. The caution indices will then be calculated and compared with the caution indices obtained by using the Sato model where the items are arranged from easiest to most difficult. It is the

intention of this study to show the use of the ordering theory to resolve the arbitrary assignment of the marginals of the items with the same p-values.

Regarding the type of group characteristics, this study will examine the characteristics of students' gender, students' socio-economic status, school location, and teachers' teaching experience.

Significance of the Study

The S-P model (Student-Problem model) is a conceptual method for identifying atypical student item responses and test items. Its intuitive appeal to educators lies in the fact that it can be quite easily interpreted. However, one of its important assumptions is the linearity of the items on the test. This assumption is plausible if the test is unidimensional. But studies have shown that many tests are multidimensional. The use of the ordering theoretic analysis of Bart and Krus (1973) to reorder the test items provides a means of constructing logical item hierarchies. This arrangement should validate the Sato model.

The other problem affecting the S-P model lies in the instability of the S and P curves due to the arbitrary assignment of marginal totals that are of the same numerical value. The ordering theoretic analysis would also resolve this arbitrary assignment of the marginal totals of the items, and serve to reduce the instability of the S and P curves. The use of the ordering theoretic analysis produces a conceptual

diagram of the hierarchical ordering of the items on the test. This would enable the educator to map the conceptual understanding of her students and thus identify their misconceptions in learning. The analysis also provides the educator with a means to identify the prerequisites of various topics being taught, and to aid in planning alternative sequences of learning experiences to cater to the needs of the students.

Knowing the impact which certain group characteristics like students' gender, school location, students' socio-economic status, and teachers' working experience, have on the caution indices would enable teachers and other concerned parties to correctly interpret the caution indices. This, in turn, would enable teachers to obtain more returns on the time and energy invested in a classroom test.

Presently, the research has shown that there exists a considerable gap between testing and instruction (Floden, Porter, Schmidt, and Freeman, 1980; Leinhardt & Seewald, 1981; Schmidt, 1983; Linn, 1983, 1990). It is hoped that this study will contribute towards narrowing this gap. There have also been various studies directed at integrating testing and instruction. Baker and Herman (1983) discussed the importance of "task structure" in integrating testing and instruction. They described "task structure" as a model of skills that are expected of the learner. Birenbaum and Shaw (1985) provided an example of the use of a task specification chart (TSC) that integrated the content facets and the procedural steps of a

specified task. They suggested that the TSC be used as tool for designing a test and for interpreting its results. Based on the cognitive theories of Piaget, Bruner, and Gagne', that there exists a hierarchical structure in learning, there exists a need to devise reliable ways in which these hierarchies may be identified and interpreted. This study, using the ordering theoretic analysis, provides a way to do this effectively. Because teachers spend as much as 12 per cent of class time for testing (Dorr-Bremme & Herman, 1986; cited in Switzer & Connell, 1990), it would be convenient and economical to employ class tests to determine these hierarchies. When these hierarchies are identified, it may be possible for teachers to identify which topics should be taught before others. In this way a closer link may be brought between testing and instruction. As Linn (1990) has aptly pointed out, "Improving the quality of classroom assessments can have a positive influence on the quality of learning." (p.425)

Research Questions

The students' responses on the test instrument were analyzed in relation to the following research questions:

1. Is there a difference between the item caution indices derived using the items ordered by the ordering theoretic analysis, and the Sato item caution indices?

2. Are the derivations of the item caution indices affected by any of the following group characteristics: format

of the test items, school location, students' gender, students' socio-economic status, and teachers' teaching experience?

3. Is there an interaction among the student caution indices between the student's socio-economic status and the school location?

Overview

It is generally accepted that a single test score on an achievement test cannot accurately measure a student's ability. Many studies have investigated students' atypical item response patterns in order to identify their misconceptions. This is in order to help bring about a closer link between testing and instruction. Several item response indices have been formulated to aid in identifying students' atypical response patterns. Among the more popular indices is the Sato caution index. The basic assumption made in using the Sato caution index is that the test being analyzed is unidimensional. As unidimensional achievement tests have been subsequently found to be multidimensional, the Sato caution index may not be accurately interpreted for all tests. Furthermore, the students-by-items matrix generated in order to calculate the caution indices employs the arbitrary assignment of marginal totals that have the same numerical value. The purpose of this study is to use the ordering theoretic analysis of Bart and Krus (1973) to construct logical item hierarchies. This hierarchical arrangement of the

items is then used in the students-by-items matrix. In this way the arbitrary assignment of those marginal totals that have the same values would be resolved. The effects of students' gender and socio-economic status, school location, and teachers' working experience on the values of the caution indices will also be examined. In doing so, it is hoped that this study will contribute towards bringing a closer link between testing and instruction.

CHAPTER II

REVIEW OF LITERATURE

Introduction

Hoge and Coladarci (1989) have stated what seems to be a widespread finding, particularly among school psychologists, educational researchers, and other professionals, that teachers are generally poor judges of the attributes of their students. This is because their perceptions are often subject to bias and error (Clark & Peterson, 1986). The increasing population of school-going children and its increasing heterogeneity are two factors that have made the public more aware of assessment in schools (Fuchs & Fuchs, 1986). This public concern has prompted related litigation (e.g. Debra P. vs Turlington, 1981) and laws (e.g. PL94-142 mandates) (both cited in Mehrens & Lehmann, 1987) to bring about a closer link between assessment and instruction.

Presently, commerically prepared standardized tests provide the most efficient and economical method to assess a large number of students. Egan and Archer (1985) explain that, "It is commonly argued that commerical tests provide teachers with valuable information about the abilities and deficiencies of their students, from which it follows that teachers who rate their students without such information will often be in error" (p. 25). This view is also shared by Mehrens and

Lehmann (1987), "... users of tests will make better decisions with appropriate data than without such data" (p. 7). They go on to elaborate that "The data provided from such tests should help teacher, counselor, administrator, student, parent, and all those concerned with teaching-learning process make the soundest educational decisions possible" (p. 10).

But, Bejar (1984) and other researchers have argued that standardized tests results, frequently have little or no impact on instruction because the test results offer little or no help in designing instruction that is optimal for an individual student. This is mainly because the standardized tests used may not be customized toward the local curricular needs. Mehrens and Lehmann (1991) make this point too when they state "Both standardized and teacher-made tests serve a common function: the assessment of the pupil's knowledge and skills at a particular time. It is usually agreed that the teacher-made achievement tests will assess specific classroom objectives more satisfactorily than standardized achievement tests" (p. 349). They add that standardized achievement-test scores may be used to supplement the empirical data obtained from teacher-made test scores to arrive at better educational decisions.

In addition to this, a single global summary score of a student's performance seldom reflects the same response pattern for another student with the same total score on the test. In other words, two students with the same score on a test may not have the same proficiency in the same content

areas of the test. Harnisch (1983) pointed out that there are as many as 252 different item response patterns that yield the same number correct score of five in a 10-item test (p.191). But he went on the say that even though 252 distinct item response patterns can be identified, it is obviously not feasible to provide different interpretations for each unique response pattern.

Interest in this area of student response patterns on a test has led to the development of powerful techniques for examining item response patterns. At least 20 different item response indices have been developed for identifying atypical response patterns. Harnisch and Linn (1981) have categorized these indices into two major groups. The first group of "appropriateness" indices is based on item response theory (IRT), while the second group is based directly on the observable pattern of right and wrong answers together with summary statistics.

The first group of item response indices based on IRT was described by Levine and Rubin (1979) with various modifications initially suggested by Drasgow (1978). Later, Wright (1979) described another example of an IRT based index of the chi-square test for person fit that is sometimes used with applications of the Rasch model. Tomic (1987) stated that at that time there were nine such different indices based on IRT. However, she classified them somewhat differently. She grouped the appropriateness measures as those which use the maximum likelihood function to estimate the item and the

student's ability. Her second group of indices were those which made use of standardized residuals to calculate the weighted or unweighted total fit mean square.

Harnisch and Linn's second group of indices include the Personal Biserial Correlation (r_i^*) (Donlon & Fischer, 1968); the van der Flier Index (U_i) (van der Flier, 1977); the Sato Caution Index (c_i) (Sato, 1975); the Dependability Indices (θ_i, θ_{ci}) (Kane & Brennan, 1980); the Agreement (A_i) and Disagreement (D_i) Indices (Brennan, 1980); the Personal Point-Biserial Index (r_i) (Brennan, 1980; cited in Harnisch & Linn, 1981); the Modified Caution Index (c_i^*) (Harnisch & Linn, 1981); the Norm-Conformity Index (NCI_i) (Tatsuoka & Tatsuoka, 1982); the Individual Consistency Index (Tatsuoka & Tatsuoka, 1982, 1983), and lastly the Person Average R (PAR) (Ayabe & Heim, 1988, cited in Shishido, Ayabe, & Heim, 1988). Harnisch and Linn also termed this group of indices as group dependent indices because they indicated whether an individual student's pattern of responses was atypical relative to the whole group of students who took the test.

Tomic (1987) went on to describe another category of indices as mathematical extensions which connect Sato's indices to the IRT model. Such extensions have been shown by Tatsuoka and Linn (1983) when they extended the concepts of Sato's Student-Problem (S-P) curve theory to take advantage of the results of item response theory. They showed analogous relationships between S-curves and the test response curves (TRC), and between P-curves and the group response curves

(GRC) in logistic models. Tatsuoka (1984) has also described the applications of such extended caution indices.

Recently, Harnisch and Jenkins (1990) developed a computer program entitled the S.P.P. (Student Problem Package) for the analysis of atypical student responses and atypical item functioning based on the Modified Caution Index. Despite the research done on the development of these group dependent indices, their practical applications have been limited. The only apparent exception is the Sato Caution Index which is widely used in Japan (Fujita, Satoh, & Nagaoka, 1977; Sato & Kurata, 1977; Tatsuoka, 1984; McArthur, 1987).

The Sato Caution Index

In 1963, Takahiro Sato, an engineer at the Nippon Electric Company (NEC) in Japan developed an instrument known as the Response Analyzed (RA) (Tatsuoka, 1978). This instrument was used to provide a teacher in the classroom with a means to determine the mean class performance and the mean response time of her pupils to a test of multiple-choice questions. Interest in analyzing class performance data with computer applications led Sato and Professor Hiroichi Fujita of Keio University to the development of a new non-parametric method of data analysis. This eventually lead to the building of the "S-P Chart" (Student-Problem Chart). Sato (1990) claims that "For the analysis of performance data, the S-P chart is preferable to the traditional test-score theory with its reliance on the normal (Gaussian) distribution of errors of

measure or the more recent latent trait theory" (p.135).

The S-P (Student-Problem) technique of analyzing patterns of student responses on a test involves the construction of a students-by-items matrix where the columns consist of the test items arranged in order of increasing difficulty, and the rows consist of students arranged from those who scored highest to lowest on the test. Sato's Caution Index, $C(S_i)$, for student i is then calculated as follows:

$$C(S_i) = \frac{\sum_{j=1}^{X_i} (1-x_{ij}) Y_j - \sum_{j=X_i+1}^{n} x_{ij} Y_j}{\sum_{j=1}^{X_i} Y_j - X_i \mu'}$$

where

x_{ij} = the i-th student's score on the j-th item, coded 1 for correct and 0 for incorrect
X_i = the i-th student's total score on the test
Y_j = the number of students getting the j-th item correct
μ' = the average item score on the test
n = the number of items on the test

Similarly, the Sato Caution Index, $C(P_j)$, for the j-th item can be calculated by using the equation below:

$$C(P_j) = \frac{\sum_{i=1}^{Y_j} (1-x_{ij}) X_i - \sum_{i=Y_j+1}^{N} x_{ij} X_i}{\sum_{i=1}^{Y_j} X_i - Y_j \mu}$$

where

μ = the average of all the students' test scores.
N = the number of students

Sato (1980) has also given the equations for both caution indices in terms of covariances.

In interpreting the caution indices, Sato (1980) used a value .5 on the caution indices for classifying students into any one of six different categories. This is summarized in Figure 2.1, reproduced from Sato (1980), Figure 8, p.155.

Similarly, using the same value of .5 in the item caution indices, Sato classified items into four categories. (See Figure 2.2, reproduced from Sato (1980), Figure 9, p.157.)

Because the Sato Caution Indices range from zero to values above 1, Harnisch and Linn (1981) have developed a Modified Caution Index, c_i^*, to produce an index with a lower bound of zero and an upper bound of 1.

The advantage of this Modified Caution Index is that it eliminates the upper extreme scores that are obtained when using the Sato Caution Index. It also serves as a basis of relative comparisons between indices. Harnisch (1983) adopted the lower value of .3 to categorize students and items with the Modified Caution Indices.

To measure the degree of discrepancy of the Student and the Problem curves to each other or to the Guttman scale, Sato developed a measure which he termed the Disparity coefficient, D.

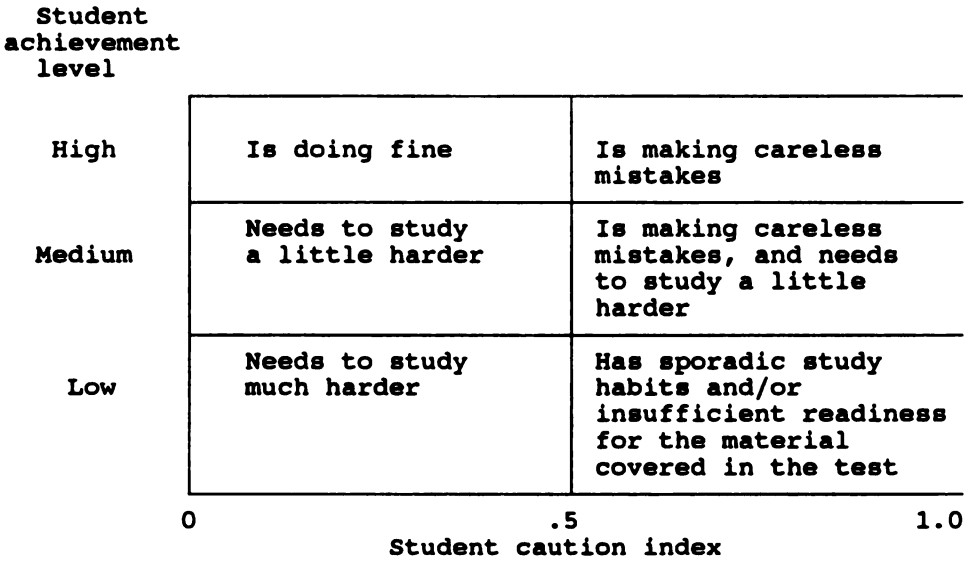


Figure 2.1 A summary of the student characteristics as reflected by their achievement level and their caution index

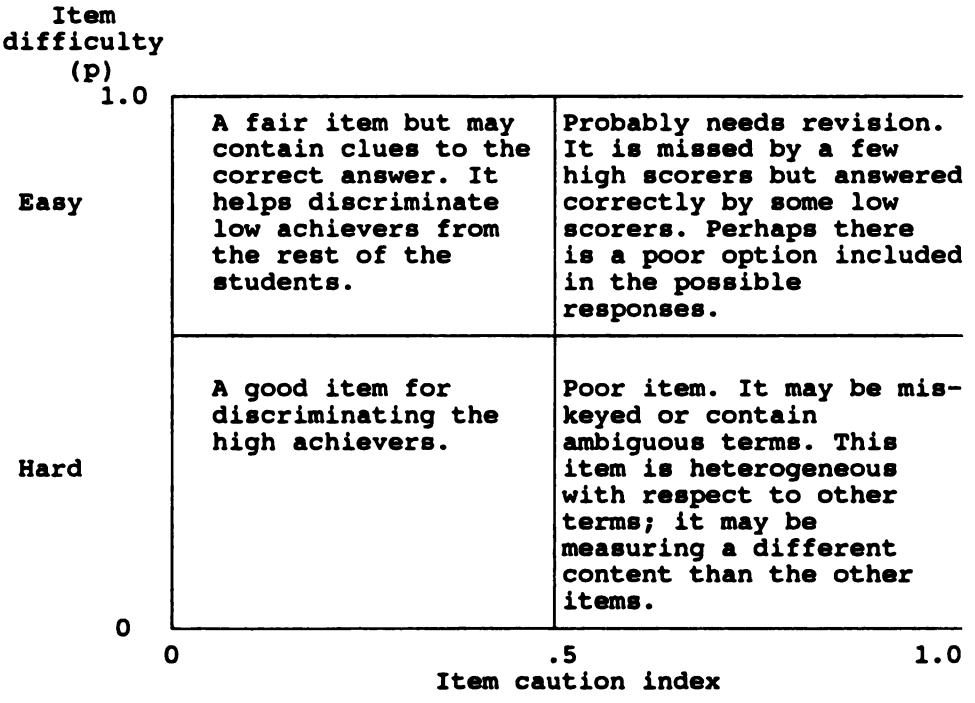


Figure 2.2 A summary of the item characteristics as reflected by the item difficulty and the item caution index

The Disparity coefficient is calculated as follows:

$$D = \frac{A(N, n, \bar{p})}{A_B(N, n, \bar{p})}$$

where

$A(N,n,\bar{p})$ is the area between the S-curve and the P-curve in the given S-P chart for a group of N students who took the n-problem test and obtained an average problem passing rate $\bar{p}$, and

$A_B(N,n,\bar{p})$ is the area between the two curves as modeled by cumulative binomial distributions with parameters N, n and $\bar{p}$, respectively.

Based on the experience of a large number of S-P charts, Sato developed a "rule of thumb" that a disparity coefficient value for an achievement test is usually around .45 to .50. This is "about right" for an ability test involving several distinct abilities (or factors), while a value exceeding .60 is a danger signal (Fujita, Satoh, & Nagaoka, 1977). In the latter case, he explains that it may signify that the set of items is excessively heterogeneous or that the group of students of two or more subgroups are having varying degrees of exposure to the material being tested.

Sato has also developed a modification of the Caution Index in order to examine patterns of responses to clusters or subtest scores in comparison with an "ideal" pattern of scores of individual subtests, namely the perfect Guttman pattern (McArthur, 1987).

Assumptions of the S-P model

All items must be scored dichotomously and students must answer all questions on the test. Any missing values where students have omitted or have not attempted the questions in the test, need to be meaningfully scored, usually with a zero.

The S-P model can be applied to as few as two students and two problems. This is to say that the model can work for a 2 x 2 students-by-items data matrix, or a 2 x J matrix, or a I x 2 matrix, where J represents the columns of the items and I represents the rows of students.

There is theoretically no upper limit for the number of students or items to which this model may be applied. The only possible limit imposed would be that of available computer memory space.

When the rows of the matrix are ordered by the total scores of the students, two or more students can share the same total score. As the positions occupied by the students must be unique, these marginal total ties must be resolved arbitrarily. Similarly, there may be ties in the item difficulty values. The resolution of this latter problem would depend on the test builder's experience and knowledge of the content material. Usually, this would be less arbitrary than in the case of students. These two steps to resolve the marginal scores would result in some instability in the S and P curves.

The calculation of the caution index depends on the linear interpretation of steps between the marginal totals.

This necessitates treating all the elements in the matrix equally and does not consider the influence of students guessing on the test.

S-P Model Interpretations

A large value of the Sato Caution Index would indicate that an atypical response pattern is present. Harnisch and Linn (1981) suggest that some of the reasons for this may be guessing, carelessness, high anxiety, an unusual instructional history or other experiential background, a localized misunderstanding that influences responses to a subset of items, or copying a neighbor's answers to certain questions. As such, they add that a large value of the caution index would raise doubts about the validity of the usual interpretations of the total score for an individual.

Harnisch's (1983) interpretations of the caution indices are referred to as Modified Caution Signals. The latter are very similar to Sato's interpretation except that Harnisch places 0.3 as the cut-off criterion value for an aberrant item. Harnisch also refers to each student's classification in terms of test performance (high or low) and the Modified Caution Index, MCI, (high or low). His four classifications for students are as follows:

Signal A = high test performance (greater than 50% of items correctly answered) and low MCI (less than or equal to 0.3),

Signal B = high test performance (greater than 50% of items correctly answered) and high MCI (greater than 0.3),

Signal C = low test performance (less than or equal to 50% items correctly answered) and low MCI (less than 0.3), and

Signal D = low test performance (less than or equal to 50% items correctly answered) and high MCI (greater than 0.3).

Similarly, his four Modified Caution Signals for items are as follows:

Signal W = difficult item (50% or fewer students answered correctly) and low MCI (less than or equal to .3)

Signal X = difficult item (50% or fewer students answered correctly) and high MCI (greater than .3)

Signal Y = easy item (greater than 50% of students answered correctly) and low MCI (less than or equal to .3)

Signal Z = easy item (greater than 50% of students answered correctly) and high MCI (greater than .3)

Precision of the Sato Caution Index

All item responses are taken to be equally meaningful. As such, the S-P analysis gives an indication of how good any single response really is. Sato (1980) suggested that a caution index value of above 0.5 would indicate the existence of an anomaly in the response pattern. Tatsuoka (1984) has suggested an index of 0.8 instead. However, Harnisch (1983) has suggested the division point of 0.3 on his Modified Caution Indices (MCIs) as indicative of atypical response patterns. He qualifies this as saying that not enough is known about the distributions of the MCIs to say that 0.3 is always a reasonable cutting point. Perhaps it would be more reasonable to leave that decision to the subject teacher as he or she may be more familiar with each of the students'

academic ability as well as the quality of the test questions built.

Advantages of the S-P Model

An advantage of using this model is that there are few assumptions made by this model. The interpretations of the model do not require a strong theoretical background. As such, most teachers, school administrators, and parents would not have much difficulty understanding the interpretation of the model.

The Sato caution indices are also reported to be less demanding to calculate than other item response indices like the Cliff's c_i1 and c_i2 indices, Mokken's H_i^* index, Tatsuoka and Tatsuoka's Norm Conformity Index (NCI) and the van der Flier's U' index (Harnisch & Linn, 1981). Harnisch and Linn (1981) have also shown that the caution index compares well to all of the indices previously mentioned.

According to McArthur (1987), the S-P technique is being mostly used in Japanese schools. An appropriate microcomputer (marketed only in Japan) has been configured exclusively for the purposes of the S-P method. This microcomputer has enabled classroom teachers to use this technique interactively (McArthur, 1987). In the U.S., there have been some efforts by Harnisch and Romy (1985) and, more recently, by Harnisch and Jenkins (1990) to apply the S-P model in a computer program. This program is presently marketed as the Student Problem Package (S.P.P.) (version 2.2) for IBM conpatible computers.

McArthur (1982) and Jaeger (1988) have also shown that the Sato caution indices may be adapted to examine various aspects of test bias.

Limitations of the S-P Model

A significant limitation of the Sato caution index is that it is group dependent since it is calculated from a students-by-items matrix. However, this is a limitation also found in many traditional psychometric analyses.

Another limitation is that the development of the S-P technique was not based on any strong psychometric or educational theory. As such, this does not allow one to draw strong inferences from the S-P model about the way in which students are performing or the manner in which the items on the test are functioning. As McArthur (1987) aptly points out, "... in developing a diagnostic interpretation of a student's score pattern, the teacher or researcher must make a conscious effort to balance the evidence in light of uncertainty about what constitutes critical or significant departure from the expected" (p.90).

Another concern is the absence of established criteria for determining the significance of the caution indices calculated. As Harnisch (1983) points out, the statistical properties and standard errors of these indices are not well understood. Little is also known about the stability of the indices when students or items having the same marginal totals are arbitrarily fixed.

Controversies about the S-P model

The S-P technique does not account for students guessing on the test. Guessing will affect the pattern of pupil responses over the items and this pattern will in turn affect the derivation of the item caution indices. As such, the interpretations of the caution indices may be potentially misleading or inappropriate.

The ordering of items according to their difficulty levels assumes linearity and unidimensionality of the data. Thus, data that are nonlinear or multidimensional will not be appropriately analyzed by the S-P method.

The Ordering Theory

Order Analysis and Dimensionality

Conceptually and intuitively, a linear ordering among a set of items represents the most parsimonious scaling of the items (Airasian & Bart, 1973). This is true if the test is unidimensional. Guttman (1950) was the first to attempt a linear ordering of the items of a test in his Scalogram Analysis. Birenbaum and Tatsuoka (1982), in their article on the dimensionality of achievement test data, stated that studies regarding the dimensionality of achievement test data in different subject matter areas have indicated that there is always more than one major factor underlying any test data in the achievement domain. Tatsuoka & Birenbaum (1979, 1981) and Birenbaum (1981) have found this result particularly for problem-solving tests and even when measuring achievement in

a specific topic.

Kingsbury and Weiss (1979) (cited in Birenbaum & Tatsuoka, 1982) have further shown, by factor analysis, that the dimensionality of a test can change depending on when the test was given. In their study, they found that the variance accounted for by the first factor at the time of the pretest was much less than at the peak of instruction. Thus, in a test on two groups of students exposed to different curricular emphases, different dimensionalities for the same test may exist. Reckase (1979) also raised such concerns about the multidimensionality of test data when applying the assumption of unidimensionality in latent-trait models.

Airasian (1971) and Airasian & Bart (1973) have shown that orderings developed from logical and statistical analyses indicate that non-linear orderings among tasks are the rule rather than the exception. This finding is supported by research in areas of cognitive development (Bart & Airasian, 1974; Airasian, Bart, & Greaney, 1975;) and curriculum development (Resnick, 1976; Gagne', 1985). Furthermore, using the Guttman scales, which measured only the linear hierarchy of tasks, made it difficult to obtain reproducible scales when there were more than six or seven tasks (Airasian & Bart, 1975).

The ordering theory developed by Airasian and Bart from tree theory, is based on scalogram analysis. It extends the analysis from linear to include non-linear hierarchical networks of tasks. Airasian and Bart (1975) define ordering

theory as "... a deterministic measurement model which uses task response patterns to identify both linear and non-linear qualitative, prerequisite relations, among tasks and behaviors" (p. 166). Its primary purpose is either to test the hypothesized hierarchies among items or to determine the hierarchies among the items (Bart & Krus, 1973).

The primary concern of the ordering analysis model is the prerequisite relationship between tasks. In the case of achievement tests, this is mapped by the test items. To apply the ordering theory to achievement tests, the items must be dichotomously scored and that the examinees must respond to all the items on the test. In the ordering analysis, one is interested in using the observed order or dominance relations between persons and items within the same set (Wise, 1981). Essentially, this means that if a person passes item i, she is said to dominate that item. Likewise, if she is able to pass item i but fails item j, then, for her, item j dominates item i.

Krus (1974) describes assymetry, transitivity, and connectedness as the essential properties of an order relation. He adds that these are the properties that provide for an inference of dimensionality of the data matrices. In terms of dominance or ordering, Wise (1981) describes asymmetry being present when elements i and j cannot dominate each other. Connectedness is explained as the existence of a relationship between two elements i and j within an order, which is to say that either i dominates j or j dominates i.

Transitivity, on the other hand is interpreted to mean that for any of the elements i, j, and k, that are in an order, if i dominates j, and j dominates k, then i must dominate k. It is this property of transitivity that permits the determination of item-item and person-person dominance in order analysis.

Thus, considering the four response patterns for two items i and j; (0,0), (1,0), (0,1) and (1,1), the item i is a prerequisite to item j to the extent that the response pattern (0,1) occurs infrequently. In terms of dominance or ordering, the response pattern (0,1) is termed disconfirmatory, and the response patterns (0,0), (1,0) and (1,1) are termed confirmatory (Bart & Krus, 1973).

The ordering analysis will provide information concerning the following types of relationships:

1. Empirical Prerequisite - A task i is determined to be empirically prerequisite to another task j if a score 0 for task i does not co-occur with a score of 1 for task j.
2. Empirically Equivalence - Two tasks, i and j, are considered to be empirically equivalent if the scores on task i are identical to scores on task j for all response patterns.
3. Empirical Independence - A task i is empirically independent of another task j if the score for task i is unrelated to the score of task j (Bart & Airasian, 1974).

Steven Wise (1986) simplifies the characteristics of these relationships in a diagram. (See Figure 2.3) However, he

prefers to use the term 'logical' instead of 'empirical' in the last two relationships shown below.

Prerequisite Relation

Task i \ Task j	0	1
0	30	0
1	50	20

Logical Equivalence Relation

Task i \ Task j	0	1
0	40	0
1	0	60

Logical Independence Relation

Task i \ Task j	0	1
0	10	20
1	40	30

Figure 2.3 Response patterns characteristic of prerequisite, logical equivalence, and logical independence relationships in percentages (Reproduced from Wise, 1986, Figure 1, p.444).

In identifying the presence of any one of these three relationships among tasks which in tests are represented by items, there is the problem of measurement error. As Wise (1986) points out, this is "Because the tasks will not be perfectly reliable measures of the model's components ... " (p.443). Therefore, the perfect response patterns of zeros in the disconfirmatory response (0,1) boxes in Figure 2.3 will rarely occur in practice. Often instead there will be a small

number of persons showing a disconfirmatory response which may be the result of random measurement errors of the items.

The most common solution to this problem is to accept a certain tolerance level for the percentage of disconfirmatory responses allowed. This value of the tolerance level should be based on the researcher's judgement of the amount of measurement error present in the data. Most often the tolerance level is set between 5% to 12% (Piazza & Wise, 1988; Wise, 1986). If the percentage of disconfirmatory responses is lower than the tolerance level chosen, then a prerequisite relationship is said to exist. Similarly, the logical equivalence relationship is said to exist when the percentage of response patterns of (0,1) and (1,0) are lower than the tolerance level chosen.

Krus (1977) also developed a probabilistic order-analytic model based on the deterministic model. The probabilistic model generates "order loadings" for the items on each dimension. For a given dimension, the order loadings reflect the relative order position of each item (Wise, 1981). Wise (1981) has suggested a modified order-analysis procedure (ORDO) which is also based on the deterministic model of Krus and Bart (1973). It serves to eliminate problems associated with item dominance and proximity by considering the partial order model of dimensionality. Essentially, it is equivalent to performing a factor analysis followed by the ordering analysis.

Reliability and Test Validity of the Ordering Procedure

Traditionally, the reliability of a test is dependent on the assumption that the variable measured by a test is unidimensional and the quality of the test is unifactorial (Bart, 1974). This implies that the variable measured is linearly ordered and the item response patterns comply to a high degree with the Guttman scalogram scale. This is to say that there is a one-to-one correspondence between summative scores and item response patterns.

According to Bart (1974), a test is reliable from an ordering-theoretic framework to the extent to which observed item response patterns conform to the true item response patterns where each true response pattern indicates the true item scores of a subject in the test. He elaborates that the test would have reliability to the extent that the item ordering of the test at time t1 was the same as that at time t2 when the same subjects were used in both cases. Alternatively, the test would have reliability to the extent that the ordering of a parallel form of a test produces the same ordering as the first test, given that a common group of subjects was used.

For traditional forms of validity, the correlational procedures are designed to measure the degree of linear relationship between linearly ordered variables. As such, predictive validity may not be quite applicable for non-linearly ordered sets of test items. However, content and construct validity would be applicable to both linearly and

non-linearly ordered sets of items.

Factor Analysis and Order Analysis

In order to use factor analysis, the data must be interval in nature, whereas order analysis only requires that the data be at least ordinal. This implies that order analysis would be more appropriate than factor analysis for evaluating the dimensionality of ordinal data. However, studies have shown differing results when order analysis and factor analysis were both used to evaluate the dimensionality of the same sets of dichotomous items. Krus and Weiss (1976), Krus (1977), and Bart (1978) found that the item orders obtained using the probabilistic order-analytic model corresponded only slightly with the factors obtained from a factor analysis. Similarly, Krus and Bart (1974), Reynolds (1981) and Wise (1981) found that there was little congruence between factors and item chains obtained from several deterministic order-analytic models. However, empirical studies using both factor and order analysis by Krus & Tellegen (1975) and Krus, Weidman, & Bland (1975) (both cited in Krus & Weiss, 1976) have frequently found the results of both methods in general agreement.

These conflicting results can best be explained by Krus and Weiss (1976) who state that the degree of congruence of factor and order analytic solutions appear to be jointly determined by the character of the data analyzed and by the values assigned to the tolerance level of the order analysis.

By comparing the two methods in two classic experiments, the Thurstone's "box problem" and Armstrong & Soelberg's experiment, they found that if the data structures are highly organized, or non-random, order analysis at any tolerance level and factor analysis would frequently converge. As the data became more random, they had to lower the tolerance level in order to find convergence with the factors.

Krus and Krus (1980) showed that a correlation matrix for two dichotomous items has two principal components representing proximal and hierarchical relationships between the two items. The proximal information is given by the proportion of (0,0) or (1,1) response patterns while the hierarchical or dominance information is given by the difference between the numbers of (0,1) and (1,0) response patterns. Thus, both proximity and dominance information is used in a factor analysis of correlation coefficients. However, in order analysis, only dominance information is analyzed. Wise (1983) (cited in Wise & Tatsuoka, 1986) showed that there were two major problems when using traditional order analysis procedures. Firstly, items that were similar in difficulty levels and measuring the same factor, for example, two parallel items, commonly do not show a dominance relation and are deemed to belong to different dimensions. This was also a finding of Wise (1981). Secondly, two items that have substantially disparate difficulty levels will tend to show consistent dominance relations, whether or not the two items measure the same factor. Hence, items measuring the same

factor may not also show a dominance relationship because they have similar difficulty levels. Similarly, items measuring different factors may show a dominance relationship only because of the difference of difficulty levels between the items.

The solution to these two problems may be found in the item proximity information. It follows that items that are measuring the same factor and are similar in difficulty level should also show high proximity. Alternatively, the items that measure different factors should show low proximity (Wise & Tatsuoka, 1986). Thus, Wise & Tatsuoka (1986) recommend that a factor analysis be performed first on the data, to identify which factors each item measures, followed by successive order analyses on the groups of items measuring the various factors.

Summary

The Sato caution index is relatively easy to compute and interpret. It involves constructing a S-P matrix where items are arranged from easy to difficult and students are arranged from highest scoring to lowest. This linear arrangement of test items in the S-P matrix implies that the test is unidimensional. However, many achievement tests are multidimensional and hence the interpretation of the caution indices of the students as well as the problems may not be accurate when considered from a unidimensional perspective. The ordering theoretic analysis permits the test items to be arranged in a hierarchical or non-linear manner. This requires

the construction of an item-student by item-student dominance matrix. However, as the ordering theoretic analysis only identifies dominance relationships, it is suggested that a principal components factor analysis be performed first to identify which items measure the factors. Following this, the ordering theoretic analysis may be conducted to identify the hierarchical order among items.

CHAPTER III

RESEARCH DESIGN AND PROCEDURE

Introduction

This study is comprised of five phases. The first phase was to obtain permission to conduct the study from the relevant authorities in Malaysia and at Michigan State University. Approval for this study was first obtained from the University Committee on Research Involving Human Subjects at Michigan State University. Permission was then sought from the Educational Planning and Research Division of the Malaysian Ministry of Education to conduct this study in six Malaysian schools. The researcher's approved proposal was submitted to both authorities for this purpose.

In the second phase, the test instrument used for gathering the data of this study was developed after the Malaysian Examinations Syndicate denied the researcher access to the student answer scripts for the 1991 national examination. The third phase was concerned with the selection of the sample of the study. In the fourth phase, the test instrument was administered, and in the last phase the data gathered were statistically analyzed.

Development of the Test Instrument

A multiple choice objective test paper was constructed and administered to a sample of 354 fifth form students (equivalent to U.S. 11th graders) from six schools. Forty test items were selected from five previous national examinations (1986-1990) for the subject of Biology. The test format conformed to that of the national examinations and the test items were selected in accordance to the table of specifications used by the Malaysian Examinations Syndicate. The latter examinations board is responsible for three of the four national examinations conducted annually in Malaysian schools.

The test was made up of two sections. Section I comprised 25 items and Section II comprised 15 items. Each section had a different multiple choice format. Both sections had five options to each item but, in Section II, each option was composed of a combination of multiple answers. In Section II, for example, the student chose option A if she believed that the first three statements after the item were true; alternatively if the student chose option B, then she believed that the first and the third statements were true, and similarly other alternatives were provided for options C, D and E. This format corresponded to what is known as the Item K format. Instructions were given on the top of each page and the students were familiar with this format as it conformed to the format of the national examination for that subject. The researcher was careful to include items that invoked some

higher order thinking skills from the thirteen categories that were listed for the subject in the syllabus. The table of specifications for the test is found in Appendix A, and the test instrument is found in Appendix B.

Selection of the Sample

Theoretically, the Sato caution index may be applied to cases of at least two students and two test items. For the ordering theoretic analysis, studies have been done with as few as 15 subjects (Bart & Krus, 1973) to as many as 1000 subjects (Bart, 1978). As for tasks, a study with as few as five Inhelder Piagetian formal operations tasks have been ordered (Bart, 1978; Bart, Frey, & Baxter, 1979;) and as many as 30 animals have been ordered in a study of the hierarchy among attitudes toward animals (Bart, 1972). However, researchers have expressed the increasing difficulty of ordering the items as the number of items increases.

The six Malaysian schools chosen for this study comprised three urban and three rural schools. The classification of these schools in terms of location conformed to that used by the authorities of the Malaysian Examinations Syndicate who were responsible for conducting pretests for the national examination. The schools chosen for the study had at least 50 candidates enrolled for the 1991 national examination in the subject of Biology. This information was obtained from the 1991 student candidature enrollment list of the Examinations Syndicate. In spite of this, when the test instrument was

eventually administered, one rural school in the study had only 36 students due to absenteeism . All of the six schools were classified by the Examinations Syndicate as having had students who were average to above average in academic performance, based on the previous students' performance in the last five years' national examinations. A summary of the number of students from each school who participated in this study is shown in Table 3.1.

Table 3.1 Information on the Subjects Used in the Study.

School location	School code	Number of students	Total number of students
Urban	School 1	79	192
	School 2	54	
	School 3	59	
Rural	School 4	59	162
	School 5	67	
	School 6	36	

Procedures of Test Administration

After selecting the schools for the study, permission was sought from the State Education Director of Selangor Darul Ehsan to conduct the Biology test, and collect demographic information on the students and Biology teachers in the six schools. On obtaining his approval, the permissions of the six principals of the schools were sought and arrangements were made to administer the Biology test to two classes of students in each school. The administration of the test to the six schools was conducted over a two-week period. All of the

schools were in the midst of preparing for their school trial examinations and most of them had one topic or less left in the syllabus to cover. This showed that all the students were approximately at the same level of preparation for the subject. This is important as comparisons of the student and item caution indices will be made between the students of different school location, gender and SES. The data should not be biased by unequal student readiness for the test.

The researcher personally administered the test to all the 12 classes of students. The purpose of the test was made clear to the students and they were instructed to answer every item on the test. The students were told that the test was a means of collecting test data by the researcher for his doctoral dissertation. Each student was given an answer sheet on which to shade their answers. This answer sheet required them to write their name, their sex and class, and the name of their school. The students were closely proctored to ensure that there was no cheating on the test. The students were given a maximum time of 75 minutes for the test. This time limit conformed to the time allocated for the Biology test in the national examinations. Almost all the students had no difficulty completing the test in this allocated time. A wall clock was placed in clear view to help the students keep track of the time. A sheet of blank paper was also given to each student for any rough calculations that they wished to make. The students were told to treat the test seriously and to answer all questions. Some students expressed the wish that

the results be made known to them as soon as possible. They were reassured that the results of this test had no bearing on their forthcoming performance in the national examinations. The data collected were meant strictly for research purposes. After the administration of the test, copies of the test instrument were given to the Biology teachers and the students were told that, if they so wished, they could discuss the questions with their teachers.

Each of the Biology teachers who taught the students in the study were given a questionnaire to fill out. When they returned the questionnaire, they were asked if any of the questions on the questionnaire needed clarification. A copy of the questionnaire is given in Appendix D. Demographic information on the students was also obtained from the school authorities. This information pertained to the individual students' school attendance and their parents' occupations. The researcher was only able to obtain the students' school attendance for the school year of 1991 as the previous year's records had been sent for auditing.

Data Analysis

The researcher was able to obtain some descriptive statistics of the previous national examinations regarding the items chosen for the test. (See Table 3.2)

The keys to the items used in the test were obtained from the Malaysian Examinations Syndicate. The student answer sheets were first hand scored by the researcher and the

Table 3.2 National Examinations Statistics for the Biology Paper for 1986 to 1990

Year	N	Items	Mean	SD	KR_{20}	SEM	M_{dcltn}	MR_{pbis}
1986	44,726	40	27.185	6.445	.831	2.650	10.990	.359
1987	49,555	40	25.770	6.980	.842	2.774	11.442	.372
1988	50,744	40	25.094	6.546	.826	2.731	11.585	.356
1989	30,687	40	23.153	6.544	.816	2.807	12.100	.349
1990	43,258	40	24.917	6.765	.842	2.689	11.610	.371

subtotals for Section I, Section II, and the combined total for both sections of the test were recorded. The students' responses and the demographic information of the students and their teachers were then coded into an ASCII (American Standard Code for Information Interchange) file. A printout of this file was obtained and each of the entries were then checked for mis-entry with each of the students' answer sheets. Following this, the file was then transferred into the SPSS program for the data analysis. The students' responses on the test were then recoded to a dichotomous score where "1" was given for a correct answer and "0" was given for a wrong answer. All missing responses were coded "9". The subtotals of the students' responses for Section I and Section II were then obtained together with the students' total score on the test. These totals were then compared with the totals obtained from hand scoring each of the students' answer sheets and found to be identical in all respects.

In analyzing the data, a preliminary item analysis was carried out and the p-values of the test items were obtained. These p-values were then compared to the p-values obtained the

same items on the national exams. A t-test was conducted for the two sets of p-values. As the test instrument was a composite of questions taken from five previous national examinations, the t-test between the two sets of p-values was to ascertain that the items on the test instrument were functioning in a similar way to that on the national tests. This is mainly because all test items in national examinations are not secure items. Histograms for the students' scores were also plotted to examine their distributions. A reliability analysis was conducted to report the Cronbach alpha coefficients for the whole test as well as for the two sections of the test.

A S-P table was then constructed with items arranged from easiest to most difficult, and students from highest scoring to lowest scoring. Prior to the calculation of the caution indices, the statistical commands of SPSS used for the analyses were tried out on the raw test scores found in a paper by Sato (1985). The caution indices reproduced for his study of 30 students and 31 problems were identical to the values he had obtained in his study. The Item and Student caution indices of the test were then calculated. The item caution indices were recalculated for various samples of the students in accordance to the type of group characteristics required by the research questions.

A principal components factor analysis was performed on the data before conducting the ordering theoretic analysis. This was to facilitate the researcher to order the items on

the test in a hierarchical fashion. A hierarchical cluster analysis was also performed on the test items. Statistical tests were then conducted on the results obtained in the analysis.

Summary

The research design of this study was aimed at addressing two main issues concerning the caution index. The first issue is the effect of the dimensionality of the test on the calculation of the caution indices and the second issue is the type of the group characteristics affecting the magnitudes of the item and student caution indices. Six Malaysian schools were selected for this purpose, three from an urban setting and the other three from a rural setting. A Biology test was constructed and administered to a total of 354 fifth form students from the six schools. Demographic information was also obtained regarding the students and the teachers who taught them the subject. All this information was coded into an ASCII file, cleaned and subjected to a data analysis. The data analysis was designed to obtain the item and student caution indices using the Sato model and the ordering theoretic analysis. A factor analysis and a cluster analysis were conducted to aid the latter analysis. Various other statistical tests were also performed on the results of the test data.

CHAPTER IV

ANALYSIS AND INTERPRETATION OF THE DATA

Introduction

This chapter presents the way in which the data analyzes were conducted. A general description of the characteristics of the sample will first be presented. This will be followed by an account of the factor analyzes and the cluster analysis on the test data. The ordering theoretic analyzes procedures and the construction of the Sato model will then be described. The manner in which the item caution indices were derived will be explained and its implications on the ordering of the items discussed. Finally, the results for the three research questions of this study will be reported together with their interpretations.

Characteristics of the Sample

A total of 354 students participated in this study. They were students in the fifth Form (equivalent to 11th graders in the U.S.). There were slightly more male than female students, and more students from the urban than the rural setting.

All of the students had at least 72% school attendance, up to the time of the test administration for the academic school year of 1991. The students' socio-economic status (SES)

was coded into two categories, low and middle/high. Students whose parents held professional jobs were classified as middle/high SES and those without professional jobs were placed in the low SES category. Table 4.1 shows the distribution of students by school location, SES, and gender.

Table 4.1 Distribution of Students by School Location, SES, and Gender

School location	Low SES		Middle-High SES		Total
	Male	Female	Male	Female	
Urban	42	20	76	53	191
Rural	58	53	22	29	162
	100	73	98	82	353[a]

[a] One female student from the urban setting did not report her parent's occupation and was excluded from the table.

A total of eight Biology teachers filled out the questionnaires of the study and their teaching experience ranged from half-a-year to 20.5 years. There was only one male teacher. With the exception of the teacher with only half-a-year's teaching experience, the other seven teachers had taught the same class the previous year. All the six schools used the same textbook and had followed the same sequence of topics for classroom instruction as outlined in the textbook. Table 4.2 shows the distribution of the teachers by teaching experience, school location, and gender.

Table 4.2 Distribution of Teachers by Teaching Experience, School Location, and Gender

Teaching experience in years	Urban		Rural		Total
	Male	Female	Male	Female	
0 - 5	0	1	0	1	2
6 - 10	0	0	0	1	1
11 - 20	0	2	0	1	3
above 20	0	1	1	0	2
	0	4	1	3	8

Analysis of the Test Data

A reliability analysis was conducted on the test items and the results are shown in Table 4.3. In the analysis of variance, the F statistic for the variation between items was significant ($F=52.695$, $p<.001$). This indicated that the items have significantly different means. This finding was confirmed by the large Hotelling's T-squared statistic ($T^2=2817.7197$) which is a test for the equality of means. Its F statistic ($F=64.4717$, $p<.001$) was significant and this indicated that the hypothesis that the items have equal means in the population can be rejected.

The hypothesis that the items are additive cannot be rejected as the F statistic for nonadditivity was not significant. This was also shown by the Tukey's test statistic which had a value close to 1. The 40-item test was reasonably reliable with Cronbach's alpha at .7509.

Table 4.3 Results of a Reliability Analysis on the Items in the Test Instrument

Source of Variation	Sum of Squares	DF	Mean Square	F	Prob.
Between persons	268.9624	353	.7619		
Within Items	3003.0500	13806	.2175		
Between items	390.0633	39	10.0016	52.695	.000
Residual	2612.9867	13767	.1898		
Nonadditivity	.0145	1	.0145	.076	.782
Balance	2612.9722	13766	.1898		
TOTAL	3272.0124	14159	.2311		

Tukey estimate of power to which observations must be raised to achieve additivity = 1.0282

Hotelling's T-squared = 2817.7197 F = 64.4717
Probability = .0000

Degrees of freedom: Numerator = 39 Denominator = 315

Cronbach's Alpha = .7509 Standardized item alpha = .7565

As one of the research questions was to compare the effects of the two item formats used in the test, the Cronbach alphas for the first 25 items (first format) and last 15 items (second format) on the test were determined. The Cronbach's alpha coefficients were found to be .6592 and .5207 respectively. When corrected for length, the alpha coefficients were .7558 and .7434 respectively. This showed that both item formats had about the same reliability. The Cronbach's alpha for all 40 items of this test instrument was lower than any of the KR20s of the five national examinations from which the test items were taken. The KR20s of the

national examinations ranged from .816 to .842 (see Table 3.2).

The point biserial and the biserial correlation coefficients of the test items were also computed. This was to ascertain if the items were all performing in the same manner in the test instrument. It must be remembered that the test items were not secure questions and students would have access to these questions. Thus it was important to see if there were any large carry over effects of some items over others. These coefficients were correlated with the point biserial coefficients of the same test items derived from the national examinations. They were found to be .794 ($p<.001$) and .744 ($p<.001$) respectively. The reasonably high correlations indicated that the items of the test were functioning in a similar manner as when they were used in the national examinations.

To determine if the sample chosen was representative of a normal population, a histogram of the students' scores on the test was plotted (see Figure 4.1).

The plot showed that the distribution of the test scores was reasonably normal. For purposes of comparison, the points of a normal curve based on all valid values of the scores were superimposed on the histogram.

Figure 4.1 Histogram frequency of 40 items on the test.

Two other histograms were plotted for the students' scores on the first 25 items, and the last 15 items on the test. They were also found to be quite normally distributed (See Figures 4.2 and 4.3).

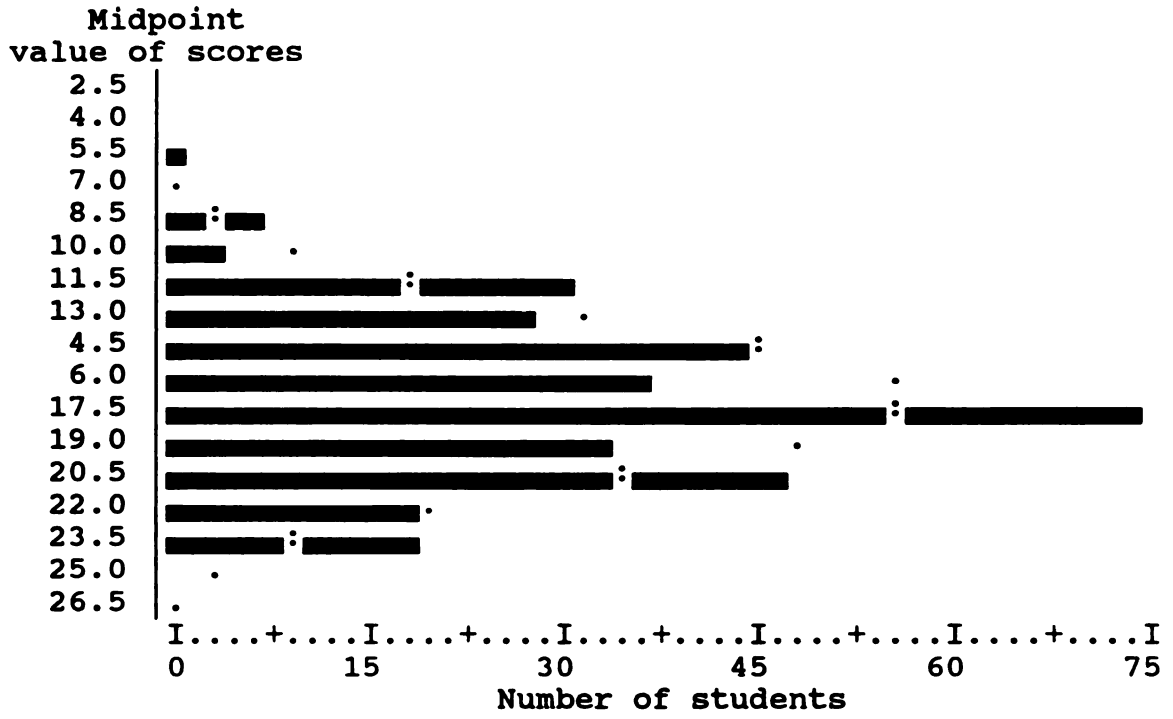


Figure 4.2 Histogram frequency of the first 25 items on the test.

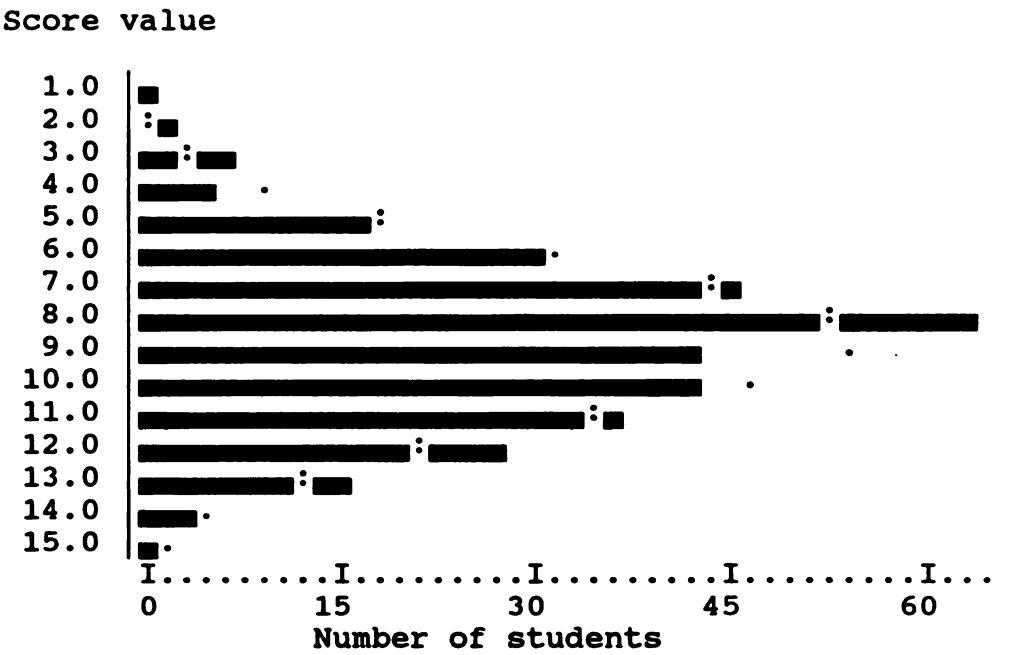


Figure 4.3 Histogram frequency of the last 15 items on the test.

A t-test was also performed on the p-values obtained from the test instrument and the p-values of the same items obtained from the national examinations. The results showed that the hypothesis that the means of the two groups were the same could not be rejected (t=-.06, df=78, p=.954).

Three principal components factor analyses were conducted on the test for all 40 items, the first 25 items, and the last 15 items respectively. The analyses showed that the factors failed to converge when the varimax rotation procedure was used. The factor analyses were repeated using the less powerful quartimax procedure. This procedure is considered less powerful because high as well as moderate factors are included in the rotation. But this procedure was adopted to enable the results to be more interpretable. The rotation redistributes the explained variance for the individual factors whereby permitting the factors to be more easily differentiated from each other. This allowed for the group of items under each factor to be identified. The subject matter of the items in each factor were then examined for any common curricular emphasis among them.

The factor analysis for 40 items on the test extracted 16 factors. The analysis for the first 25 and last 15 items on the test extracted 10 and 6 factors respectively.

The results of the factor analysis with 40 items showed that the Bartlett's test of sphericity was large (1635.3703). This indicated that it was unlikely that the population correlation matrix was an identity. The Kaiser-Meyer-Olkin

measure of sampling adequacy at .72062 was reasonably large. These two statistics justified performing a factor analysis on the data.

The 16 factors extracted by the factor analysis accounted for 58.6% of the variance which left 41.4% of the total variance still unexplained. There were also low interitem correlations of the items that were grouped into factors in the rotated factor matrix. This finding is similar to the findings of Wise (1981), Reynolds (1981) and Krus and Bart (1974) when they found little congruence between factors and the item chains from several deterministic order-analytic models. This is also probably why the initial varimax rotation procedure was unsuccessful. Due to the low interitem correlations, the 16 factors extracted from the factor analysis were not very meaningful. An examination of the various topics represented by the items that were grouped into the factors, did not make any strong conceptual sense. Similar findings were found for the other two factor analyses.

The factor analyses were performed primarily to facilitate the interpretation of the results of the ordering analysis. However, the factors extracted were not able to do this. The reason for this may be explained by Krus and Weiss (1976) who state that the degree of congruence of factor and order analytic solutions appear to be jointly determined by the character of the data analyzed and by the values assigned to the tolerance level of the order analysis. In this case, the data may not be highly organized in a hierarchical manner.

Other reasons for the failure of the factor analysis to be more informative were largely due to the moderate inter-correlations between the items, and the fact that binary data were used in the analysis. In addition, the sample size used was too small. According to Nunnally (1978) there needs to be at least 10 subjects per item for the factor solutions to be stable. In this study there were 40 items with only 354 students.

In this case, the scree plot which is the plot of the total variance associated with each factor showed that there was essentially one factor. Again it may be deduced that perhaps the data may not be highly organized in a hierarchical fashion.

An agglomerative hierarchical cluster analysis using the complete linkage method was performed on the data. Six clusters could be identified from the dendrogram produced. It was interesting to note that the ordering of the items produced by the complete linkage method was very similar to the order of items of the Sato model. A Spearman rank correlation between the order of items produced by the cluster analysis and the Sato model produced a $r_{ranks} = .903$. The dendrogram of the cluster analysis is found in Appendix E.

The Ordering Theoretic Analysis

The results of the ordering theoretic analysis were first interpreted by using a tolerance level of 10 per cent. At this

tolerance level, the researcher was unable to accurately construct a conceptual map using all 40 items of the test. A 10 per cent tolerance level meant that if the percentage frequency of the confirmatory response of a pair of items was more than 10 per cent of the total possible responses, the prerequisite relationship was considered to exist. However, using the same tolerance level with the first 25 items on the test did reduce the difficulty of the task of constructing the conceptual map. This task became much easier with the last 15 items of the test. This confirmed the findings of other studies that reported the increasing difficulty of constructing a conceptual map when the number of items to be ordered became large.

Further analysis also showed that if the tolerance level was lowered, it also became less difficult to construct the conceptual map. However, somewhat different conceptual maps were produced depending on the choice of the tolerance level. For example, using the tolerance level of 10% on the last 15 items of the test produced a conceptual map of three hierarchical levels. Lowering the tolerance level to 5% produced a conceptual map with six hierachical levels. Where in the former the lowest item in the conceptual map, item 34, was a prerequisite to ten items, in the latter it was a prerequisite to only one item. Realizing the inconsistencies produced by this deterministic method of analysis, the researcher decided to adopt Krus's (1975) solution of constructing logical item hierarchies in one dimension. In his

method, Krus recommended the use of the probabilistic model Z. He explained that deterministic models of medium-sized data matrices "... frequently yield structures which are too complex and difficult to interpret" (p. 56, cited in Krus, Bart and Airasian, 1975). The probabilistic model uses the McNemar's z criterion score for nonindependent proportions instead of a tolerance level. McNemar's z criterion score is given by the formula:

$$z=\frac{d_{ij}-d_{ji}}{\sqrt{d_{ij}+d_{ji}}}; \quad i\neq j$$

where d_{ij}= confirmatory responses (1,0) between item i and j

d_{ji}= disconfirmatory responses (0,1) between item i and j

The z criterion scores were calculated for all possible pairs of items in the test. Usually, z criterion values are selected with the accepted rules of significance testing where 1.96 is the critical value for 5% level for two-tailed tests. For a particular pair of items, positive values of 1.96 and above indicated that item i served as a prerequisite to item j. This meant that the item j was above item i in the conceptual map. Conversely, if the z criterion value was -1.96 and below, item j is a prerequisite to item i, and is placed below item i in the conceptual map. The number of positive and negative significant values was then recorded. These results were then tabulated in descending magnitude of the number of items that a particular item is a prerequisite to (see Table

4.5). However, there were instances when items had the same number of prerequisites. Ties in the position of the test items were resolved by going back to the particular items in question and selecting the item with the next z value closest to the absolute value of 1.96. The item with the closest negative value would be placed highest in the order of items in the table. Conversely, the item with closest positive value would be placed lower in the order. This gave the sequence of ordering of the test items as shown in Table 4.5.

This arrangement of the items gave the sequence in which topics represented by the items in the test were ordered by prerequisite conditions.

Ordering the Items According to the Sato Model

To begin the analysis using the Sato model, a students-by-items matrix was constructed where the 40 items were ordered from easiest to most difficult and the students from highest scoring to lowest scoring. To verify if the SPSS/PC+ commands used for this analysis were correct, the commands were first applied to the test data found in an article by Sato (1985). The caution indices computed for both the students and the items were found to be identical to those obtained by Sato in the mentioned article. The same commands were then applied to the study's data to compute the caution indices.

Table 4.5 A Distribution of the Test Items According to Prerequisite Requirements.

Order of test item	Test item number	No. of items prerequisite to this item (1.96 and above)	No. of items this item is a prerequisite to (-1.96 and below)
1	4	0	35
2	12	0	34
3	34	0	34
4	18	0	34
5	5	0	33
6	10	1	33
7	3	4	27
8	8	6	25
9	1	6	24
10	6	6	24
11	32	6	24
12	9	6	23
13	11	6	23
14	16	7	23
15	19	7	23
16	25	7	23
17	31	11	17
18	36	16	14
19	35	15	13
20	37	15	12
21	29	15	12
22	13	15	12
23	20	16	12
24	39	15	10
25	28	17	8
26	17	17	8
27	40	18	8
28	21	19	6
29	15	23	6
30	22	23	6
31	2	24	6
32	26	24	6
33	14	27	5
34	7	27	4
35	33	32	4
36	24	33	4
37	38	36	1
38	27	36	1
39	30	36	0
40	23	38	0

Answers to the Research Questions

Three research questions were formulated in this study. In the following pages, each research question will be stated and the results of the analysis pertaining to that question will be reported.

Research Question 1

> Is there a difference between the item caution indices derived using the items ordered by the ordering theoretic analysis, and the Sato item caution indices?

The analysis showed that the items arranged according to the Sato model and according to the ordering theory model produced identical values for the item caution indices. This implied that the ordering of the items had no effect on the item caution indices. On closer investigation, it was noticed that Sato's formula used for the computation of the item caution indices did not consider the arrangement of the items in the students-by-items matrix. The item caution indices were calculated using the simplified formula below (Sato, 1980):

$$C(P_j) = 1 - \frac{\frac{1}{N}\left(\sum_{i=1}^{N} x_{ij}X_i - Y_j\mu\right)}{\frac{1}{N}\left(\sum_{i=1}^{Y_j} X_i - Y_j\mu\right)}$$

where

- $C(P_j)$ = the caution index for the j-th item
- x_{ij} = the i-th student's score on the j-th item, coded 1 for correct and 0 for incorrect
- X_i = the i-th student's total score on the test
- Y_j = the number of students getting the j-th item correct
- μ = the average of the students' test scores
- N = the number of students

In order to show why the calculation of the item caution index is not influenced by the arrangement of the items in the S-P table consider the derivation of the item caution index for item P_4 in Figure 4.4.

Problems (Items)

students	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	Total score
1	1	1	1	1	1	1	1	1	1	1	10
2	1	1	1	1	1	1	1	1	1	0	9
3	1	1	1	1	1	0	1	1	0	1	8
4	1	1	1	1	1	1	1	0	0	0	7
5	1	1	1	1	1	0	1	1	0	0	6
6	1	1	0	0	1	0	1	0	1	0	6
7	1	1	1	1	0	1	0	0	0	0	5
8	1	1	1	0	1	0	0	0	1	0	5
9	1	1	0	1	0	0	1	0	0	1	5
10	1	0	0	1	0	1	0	1	1	0	5
11	0	1	1	1	0	1	0	0	0	0	4
12	1	0	0	0	0	1	0	1	0	1	4
13	1	0	1	0	1	0	0	0	0	0	3
14	0	0	1	0	0	1	0	0	0	0	2
15	0	1	0	0	0	0	0	0	0	0	1
	12	11	10	9	8	8	7	6	5	4	

Correct answers

Figure 4.4 An ordered students-by-items matrix (S-P table) for 15 students and 10 problems

The first term in the numerator on the right side of the equation is given as follows:

$$\sum_{i=1}^{N} x_{ij} X_i = 10+9+8+7+6+0+5+0+5+5+4+0+0+0+0 = 59$$

It follows that this term in the equation is not influenced by the arrangement of the students in the S-P table. For the second term in the numerator,

$$Y_j = \sum_{i=1}^{N} x_{ij} = 9$$

This is the number of students getting item P_4 correct.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i = \frac{1}{15}(10+9+8+7+6+6+5+5+5+5+4+4+3+2+1) = \frac{80}{15}$$

This term gives the average student's score for all the 15 items. The first term of the denominator is as follows:

$$\sum_{i=1}^{Y_j} X_i = 10+9+8+7+6+6+5+5+5 = 61$$

This term gives the cummulative sum of the students' scores from student #1 to student #9 (i.e., the number of students getting item P_4 correct).

Combining the terms to obtain the item caution index,

$$C(P_4) = 1 - \frac{\frac{1}{15}(59 - \frac{9x80}{15})}{\frac{1}{15}(61 - \frac{9x80}{15})} = .154$$

Therefore, the arrangement of the items on the S-P table does not influence the derivation of the item caution indices.

Similarly, it can be shown that the derivation of the student caution indices are not influenced by the arrangement

of the students in the S-P table. Using the same notation, the student caution index can be derived by the following modified formula (Sato, 1980):

$$C(S_i) = 1 - \frac{\frac{1}{n}\left(\sum_{j=1}^{n} x_{ij}Y_j - X_i\mu'\right)}{\frac{1}{n}\left(\sum_{j=1}^{X_i} Y_j - X_i\mu'\right)}$$

where

$C(S_i)$ = the caution index for the i-th student
μ' = the average number of students on the test getting the problems correct
n = the number of items

The arrangement or ordering of the students in the matrix would affect only the calculation of the item caution indices. Although the ordering theory provides an alternative way to arrange the items on the test, at present, there is conceptually no other logical or statistical method of ordering the students other than from highest to lowest scoring.

The arrangement of the items based on the probabilistic model z of the ordering theoretic analysis was very similar to Sato's arrangement of easiest to most difficult item. A Spearman rank correlation conducted between the two sets of items showed the items from both models to be almost identically ranked (r_{ranks}=.999). The researcher also arranged the items according to the sequence in which all the thirteen topics of the subject were taught to the students. For example, all items pertaining to topic 1 were placed first, according to decreasing p-values, followed by topic 2 and so

on, until topic 13. This was done in order to see if there was any difference between the arrangement of the items according to the Sato model and the order of the items according to the sequence in which the topics were taught to the students. A Spearman rank correlation of the items on the test in the instructional sequence of the topics and the items arranged according to the Sato model was conducted. A low r_{ranks} of .311 was obtained. Similarly, a Spearman rank correlation of the items arranged in the instructional sequence with the items in the arrangement of the ordering theory was also low (r_{ranks}=.260).

The analysis of the test items using the Sato model produced 28 items with caution indices above the value of .50. Using Sato's classification criteria of items with caution indices above .50 as indicative of an item for review, these results were most unusual. This was because these items on the test instrument were first pretested, statistically analyzed, and found to be good before they were included in the national exams. As such they were considered to be reasonably good items that were technically sound. A perusal of their national examinations point biserials showed only three items to be between .080 and .173 while all the other 37 items were between .200 and .528 (see Appendix D). This suggests that Sato's classification criterion of .50 may be too severe for this data set. Perhaps the use of .80 for the criterion as suggested by Tatsuoka (1984) may be more appropriate. Using .80 it was found that only three items were classified as

atypical. These three items were observed to have low point biserial correlations in the test instrument, namely from .00 to .10. The mean of the point biserial coefficients for the 40 items of the test instrument was .3076.

The item caution indices were also calculated using the responses of 79 students from the school that the researcher considered to be the best performing school in the study's sample. It was found that there were now 29 items that had item caution indices above .50. This was two more items than when the total of 354 students were used to calculate the items caution indices. A t-test between the means of the item caution indices of the two samples showed that they were not significantly different from each other (t=.16, p=.874). A comparison of the values of the caution indices showed that they were only higher by a mean of .01 when the smaller sample was used. The two sets of item caution indices were also moderately correlated to each other (r= .6296).

The 13 items that had item caution indices below .50 in the study's total sample were then selected and the item caution indices recalculated using only those items. It was found that the item caution indices were all still below .50 but were a mean of .06 lower in value than when the item caution indices were calculated using the full 40 items of the test. It appeared from these findings that the value of the item caution indices derived were sensitive to the number of items.

A similar investigation was carried out for the student caution indices. The values of the student caution indices for each of the classes from three of the schools were compared when they were derived by i) using the whole sample of 354 students, ii) using the school sample only, and iii) using the class sample only. (See Table 4.6)

Table 4.6 Mean Student Caution Indices According to Sample Size and Number of Students at or above .50

Sample	Sample size	Mean student caution index	No. of students at or above .50
Test sample	354	.468	17
School 1	79	.431	16
Class 1	37	.418	10
Test sample	354	.464	16
School 1	79	.433	16
Class 2	42	.403	13
Test sample	354	.495	18
School 5	67	.436	13
Class 1	33	.382	9
Test sample	354	.624	25
School 5	67	.611	29
Class 2	34	.558	23
Test sample	354	.654	14
School 6	36	.523	10
Class 1	18	.475	8
Test sample	354	.618	14
School 6	36	.504	8
Class 2	18	.476	6

In all the three schools it was found that the mean student caution index decreased as the sample of students used in the derivation decreased. What was, perhaps more

significant was the number of students that were classified at or above a value of .50 of the student caution index, decreased drastically.

In sum, it may be said that the findings of the study report that the Sato caution indices for the students and the items are sensitive to the number of students taking the test and the number of questions on the test.

As the arrangement of the students in the students-by-items matrix determined the values of the item caution indices, the researcher decided to determine if the student caution indices could be used to arrange the students in the matrix. Using arrangement of the items of the Sato model, the student caution indices were first calculated. The students were then rearranged in ascending values of their caution indices (i.e. from lowest to highest student caution index). This new arrangement of students were then used to calculate the item caution indices. It was found that this produced 37 item caution indices with negative values. Theoretically, the Sato caution indices can only have positive values. As negative values of the caution indices were obtained, it was inferred that the formula derived by Sato for the caution index could not be applied for this arrangement of items and students. This confirms the basic assumption that the Sato model is only appropriate for linear and unidimensional data. Therefore, this direction of investigation was abandoned.

The student caution indices were also calculated after the items had been arranged according to the instructional

sequence of the topics. Again negative values of student caution indices were obtained and it was inferred that the Sato model could not be applied in this case also because of the violation of the linearity and unidimensional assumptions.

Research Question 2

> Are the deviations of the item caution indices affected by any of the following group characteristics: format of the test, school location, students' gender, students' socio-economic status, and teachers' teaching experience?

In the following account, each group characteristic was considered separately and the testing of six research hypotheses concerning these group characteristics are reported with regard to their statistical significance.

There were two formats used in the test instrument. The first 25 items (Section I) was one format, and the last 15 items (Section II) was the other. The item caution indices derived from both formats were statistically tested with the item caution indices derived from the complete test.

To answer the research question concerning the item formats, two statistical hypotheses were tested. Each hypothesis is stated, followed by the results of the statistical analyses for that hypothesis.

> Hypothesis 1:
> There is no difference between the item caution indices of the first 25 items on the test that were derived by using all 40 items on the test and that derived by using only the first 25 items on the test.

The results of a paired t-test showed that the correlation between the two sets of item caution indices was high (r=.968), and there was sufficient evidence to reject the hypothesis that the caution indices have similar mean values (t=2.92, df=24, p<.01). The mean of the item caution indices of the first 25 items of the test derived by using all 40 items on the test was a little larger (.5780) compared to that for the 40 items (.5548). This result supports earlier findings that the derivation of the item caution indices are influenced by the number of items on the test.

Hypothesis 2:
There is no difference between the item caution indices of the last 15 items on the test that were derived by using all 40 items on the test and that derived by using only the last 15 items on the test.

The results of a paired t-test showed that the correlation between the two sets of item caution indices was also high (r=.982). There was also stronger evidence in this case to reject the hypothesis that the caution indices have similar mean values (t=11.15, df=14, p<.001). The mean of the item caution indices derived by using the last 15 items was found to be .5287. The mean of the item caution indices of the same items derived by using all 40 items was found again to be larger (.6120).

These results again indicate that the calculation of the item caution indices is affected by the number of items on the test. In examining the values of the item caution indices and

using Sato's classification criteria of 0.50 and above, as an item to be considered as atypical, it was interesting to note that there were discrepancies in classification for six items. Three of the items were in the first 25 and three others in the last 15 items of the test. With the exception of one item, the other five were all lower in value for the respective subtest. A closer look revealed that the difference in values was smaller for the three items of the first 25 items than for the last 15 items. The discrepancies in the item caution indices are shown in Table 4.7.

Table 4.7 Item Caution Indices of Six Discrepant Items

Item Caution Index Derivation Method	Item Number 4	14	24	31	34	40
Using all 40 items	.47	.54	.50	.53	.52	.54
Using the first 25 items	.51	.49	.49	-	-	-
Using the last 15 items	-	-	-	.47	.44	.47

To graphically show how much difference there was between the item caution indices due to the method of derivation, a standardized measure of the differences was derived. The item caution indices derived by using only the first 25 and the last 15 items were subtracted from the item caution indices derived by using all 40 items on the test. The values obtained were then divided by the mean item caution index obtained from using all 40 items of the test. These values were then plotted against the items (See Figure 4.5). The plot showed that item caution indices calculated using the last 15 items on the test

tend to be lower in value than those item caution indices calculated using all 40 items on the test. This is indicated by the presence of all the points for the last 15 items being above zero of the vertical scale (D).

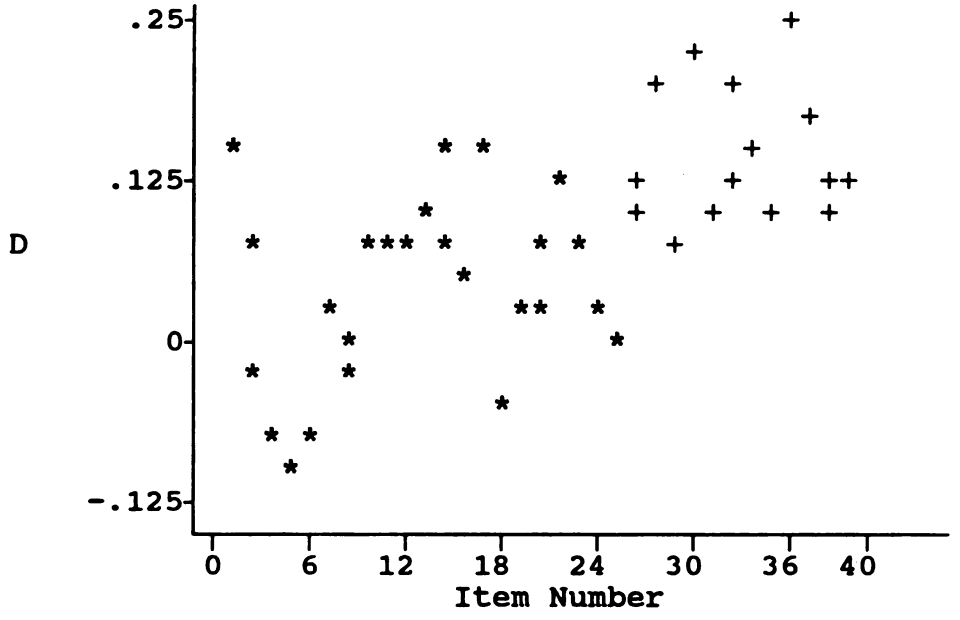


Figure 4.5 A plot of the standardized difference in item caution indices (D) with its corresponding item number (Note: * first 25 items, + last 15 items)

It was also found that using only the first 25 and the last 15 items on the test lowered each of the item caution indices by a mean of .02 and .08 respectively.

The analysis of the test data with the Sato model also showed 28 out of 40 items with the item caution indices at or above Sato's critical value of .50. When the item caution indices were examined separately under the two different formats of the test, only a combined 24 out of 40 items had item caution indices above at or above 0.5. It may be argued that this smaller number of aberrant functioning items may not

be entirely due to the number of items used for analysis because there was a difference in the type of multiple choice questions used in the two formats. However, a oneway analysis of variance between the item caution indices of the two formats showed that they were not significantly different from each other (F=.3441, p=.5610). Studies have shown that items of the K format similar to those used for the last 15 items on the test instrument have been found to have lower p-values. However, in the Sato model, the last 15 items of the test instrument were quite widely spaced out in terms of p-value (See Appendix F). A correlational analysis conducted on the item caution indices derived by using all 40 items and the composite 40 item caution indices that were derived from section I and II of the test showed that they were very closely correlated (r=.9498).

Hypothesis 3:
There is no difference between the item caution indices derived from students in the urban setting and students in the rural setting.

The t-test indicated that there was insufficient evidence to reject the hypothesis that the means of the item caution indices were the same (t=-.37, df=78, p=.709). The correlation of the item caution indices derived from the students of the different school settings was moderately large at .6611. This suggests that the location of the school has no significant effect on the derivation of the item caution indices.

Hypothesis 4:
There is no difference between the item caution indices derived from male students and female students.

The t-test showed that there was insufficient evidence to reject the hypothesis that there is no difference between the means of the item caution indices derived from the male and the female students (t=.99, df=78, p=.326). There was also a moderate correlation of .5437 between the item caution indices derived from the male and female students.

Hypothesis 5:
There is no difference between the item caution indices derived from students of low SES and students of middle/high SES.

The t-test indicated that there was insufficient evidence to reject the hypothesis that there is no difference between the means of the item caution indices derived from the two groups of students (t=1.39, df=78, p=.169). This suggests that the grouping of students according to socio-economic status has no significant effect on the derivation of the item caution indices. There was also a moderately large correlation between the item caution indices derived from the two groups of students (r=.6298).

Hypothesis 6:
There is no difference between the item caution indices derived from the teachers with 5 years or less, 10 years or less, and those with more than 10 years of teaching experience.

The one-way analysis of variance for differences in teaching experience showed that there was insufficient evidence to reject the above hypothesis (F=.4552, df=2, p=.6355).

Table 4.8 One-Way Analysis of Variance on the Differences in the Mean Item Caution Indices Derived from Teachers with Different Teaching Experience

Source of Variation	DF	Sum of Squares	Mean Squares	F Ratio	F Prob.
Between Groups	2	.0242	.0121	.4552	.6355
Within Groups	117	3.1150	.0266		
Total	119	3.1393			

This suggests that the years of teaching experience did not have a significant effect on the mean item caution indices that were derived. The correlation between the item caution indices derived from the students of the teachers with five or less years of teaching experience with those indices from teachers of more than 10 years teaching experience was low (r=.3935, p=.012). The correlation was highest with the indices of the most inexperienced teachers and the teachers with 10 or less years of teaching experience (r=.8426). There was a moderate correlation between the indices of the most experienced teachers and those with 5 or less years of teaching experience (r=.6339).

Research Question 3

> Is there an interaction among the student caution indices derived from students with different socio-economic status and the student caution indices derived from students from different school locations?
>
> The hypothesis to answer this question is:
>
> The student caution indices derived from students from different school locations does not have an effect on the student caution indices derived from students from different socio-economic status?

The two-way analysis of variance conducted to test this hypothesis showed that there was insufficient evidence to show that the student caution indices derived from different school locations did have an interaction effect on the student caution indices derived from students from different socio-economic status (SES). See Table 4.9.

Table 4.9 Two-Way Analysis of Variance on the Effect of Students' SES and the School Location

Source of Variation	Sum of Squares	DF	Mean Square	F	Sign of F
Main Effects	.541	2	.271	8.469	.000
Students' SES	.007	1	.007	.207	.649
School location	.506	1	.506	15.845	.000
2-way Interactions	.077	1	.077	2.416	.121
Students' SES & School location	.077	1	.077	2.416	.121
Explained	.618	3	.206	6.451	.000
Residual	11.150	349	.032		
Total	11.769	352	.033		

Note: 353 Cases were processed and there was 1 missing case.

It is interesting to note that the main effect of school location was significant when students' SES was held constant even though there was no significant interaction effect. Correlations between the school location and the students' SES were high, ranging from .7067 to .9103. This implies that the location of the school has some influence on the students' performance regardless of the students' SES. The student SES-by-school location matrix of the mean student caution indices is shown in Figure 4.6 below.

Student SES	School location Urban	Rural	Row mean
Low	.470 (n=62)	.580 (n=111)	.525 (n=173)
Middle/ High	.510 (n=129)	.560 (n=51)	.535 (n=180)
Column mean	.490 (n=191)	.570 (n=162)	.530 (n=353)

Figure 4.6 Student SES-by-school location matrix of student caution index means

Summary

The data of this study were the responses of 354 students to a Biology test. Preliminary steps were taken to ensure that the data were appropriately coded and sufficiently accurate before the start of the analysis. The statistical methods used in this study to derive the Sato caution indices were also verified to be correct by administering them to a set of data

with known caution indices. Initial analysis of the data showed that the data were normally distributed. The mean of the p-values was found to be not significantly different from the mean of the p-values of the items derived from the respective national examinations. A principal components analysis revealed 16 factors and did not prove to be of help in the ordering analysis. Due to the complexity of the conceptual map produced by the items, the probabilistic model Z of the ordering theory was adopted for the analysis using the z criterion of 1.96. The analysis showed that the linear arrangement of the items did not have a bearing on the item caution indices but only affected the student caution indices. Statistical tests were conducted on hypotheses formulated to answer the three research questions of this study. The item caution indices derived from the different group characteristics of item format, school location, students' gender, students' SES, and teachers' working experience, were all not significantly different. There was also no significant interaction effect between the student caution indices derived from students of different SES and different school locations. The student caution indices derived from different school locations were significantly different when students' SES was held constant.

CHAPTER V

SUMMARY, CONCLUSIONS, IMPLICATIONS, AND RECOMMENDATIONS

Summary of the Purposes and Procedures of the Study

A group of 354 Malaysian students from six secondary schools were selected for this study. Three schools were located in an urban setting and three others in a rural setting. A Biology achievement test for the 11th grade level was constructed from items selected from previous national examinations and administered to the students. The students' responses to this 40-item multiple choice objective test were analyzed and their item and student caution indices were calculated according to the Sato model. The ordering theoretic analysis was then applied to the test data and the caution indices recalculated. Three research questions were formulated for this study and tested under a total of seven hypotheses. Statistical data analysis techniques of factor analyses, t-tests and analysis of variances were employed to answer the research questions.

The purpose of this study was to investigate two main issues concerning the Sato caution index. The first issue addressed the arrangement of the items in the students-by-items matrix constructed for the purpose of calculating the caution indices. The Sato model adopted the arrangement of

items according to descending p-values. In this study, the probabilistic model z of the ordering theoretic analysis was employed to construct logical item hierarchies in one dimension. The z criterion value of 1.96 was used to order the items in the students-by-items matrix. The study compared the item caution indices derived by using these two ways of item arrangement and found them to be identical. It was also found that the arrangement of the items in the students-by-items matrix did not affect the values of the item caution indices derived. However, the arrangement of the items did affect the derivation of the student caution indices. Similarly, the arrangement of the students in the matrix affected only the derivation of the item caution indices. An alternative method of arranging the students was proposed but did not prove to be viable because it violated the assumptions of linearity and unidimensionality. It was also found that the number of students and the number of items on the test did affect the caution indices. Reducing the number of students had an effect of lowering the student caution indices, and reducing the number of items in the test had an effect of lowering the item caution indices. It was also found that reducing the number of students did not significantly affect the item caution indices.

The second issue investigated by this study was the type of group characteristics affecting the magnitudes of the item and student caution indices. The factors studied included the two item formats used in the test, two different school

locations, the gender and socio-economic status of the students, and the working experience of the teachers.

The findings of the study showed that there were no significant differences between the mean item caution indices of the two formats of the test, the different school locations, the students' gender, the students' socio-economic status and the teachers' working experience. There was a significant effect of the student caution indices for school location when holding the student SES constant.

Discussion and Conclusions

The Test Data

The analysis conducted on the students' responses showed that the data obtained were reasonably representative of the actual population. The data gathered were normally distributed and the means of the groups were not significantly different. The point biserials of the test items correlated quite well with the biserials of the same items in the national examinations (r=.7944).

The Sato Model

The results of the analyses showed that in the Sato model, the caution indices were affected by the number of students taking the test as well as the number of items on the test. Reducing the number of items on the test, lowered the mean item caution indices that were derived. Similarly, reducing the number of students on the test, lowered the mean

student caution indices that were derived.

The Ordering Theoretic Analysis

The probabilistic model z of the ordering analysis produced an ordering hierarchy of items in one dimension that was remarkably similar to that of the Sato model. The Spearman ranks correlation showed the arrangement of the items of both models to be almost identical (r_{ranks}=0.999). This logical hierarchical arrangement of items did not clearly group the items in terms of the topics they represented in the table of specifications. This may be because the topics of Biology were themselves inter-related. The factor analysis showed that these interrelations were not very strong. An investigation of the arrangement of the topics represented by the items resulting from the ordering theory did show that the more complex topics of growth, variation, reproduction, and mechanisms of coordination, were found higher up in the hierarchy. Likewise, more basic topics concerning the soil and the role of water and body fluids were found lower down in the hierarchy. The ordering theoretic analysis did indicate that the difficulty of the topics was strongly related to the difficulty of the items on the test. This provided the theoretical basis for Sato to adopt descending p-values in the students-by-items matrix. However, Sato's formula used for deriving the item caution indices, does not take into consideration the arrangement of the items in the matrix. The arrangement of the items had a bearing only on the student

caution indices. Conceptually, this appears to be sound as the student caution indices would indicate how well the logical hierarchy of items is understood by the student.

It was found that the arrangement of the students in the students-by-items matrix has a bearing on the derivation of the item caution indices. Because there is no conceptual basis for arranging the students from highest scoring to the lowest scoring, the values of the item caution indices obtained should not always be taken to mean that the item is not functioning well. This is also because there is the arbitrary assignment of the students with the same score at different positions in the matrix. If a large number of students taking the test obtain the same total score then this would have an influence on the item caution indices derived. This is especially significant if there were only a small number of items on the test. Therefore, the item caution indices should also be conceived as an indicator of the interrelationship of the topic represented by that item with the other topics in the hierarchy of the items. Thus, a high item caution index for an item may mean, among other things, that the interrelationship of that item with other items was not clearly understood by the students. However, this inference can only be accurately drawn if the teacher was reasonably confident that her test item was functioning well and other factors like guessing and cheating on the test, were held constant. In this study, reasonably reliable items were chosen from past national examinations. As such it was inferred from

the large number of items with high item caution indices, that there was a poor general understanding by the students, of the topics represented by the arrangement of items. It was also inferred that Sato's classification criterion value of .50 and above for the caution indices may not always be appropriate. It is suggested that for a teacher using this method of analysis, the classification criterion needs to be set by the teacher based on her confidence of how well the test items were constructed and how well her students were generally performing in class.

The results of the analysis suggests that the point biserial or possibly the biserial coefficients may be a more accurate than Sato's caution index of .50 as indicator of an aberrant item. If the point biserial coefficients are used, they should be interpretated in the same way Ebel (1972) used the index of discrimination for item evaluation. This means that any item with a point biserial below .19 should be viewed as an aberrant item.

Group Characteristics

Item Format

The results of the analysis on the formats of the test showed that the mean item caution indices calculated using the first 25 and the last 15 items differed significantly from the mean item caution index calculated using all 40 items on the test. This implied that the values of the item caution indices were influenced by the number of items used in a test. Hence,

an item caution index may change depending on whether it is derived from a long or a short (in terms of number of items) test. There were indications in this study to suggest that a shorter test would produce lower values of the item caution indices. As such, Sato's criterion of .50 of the caution index of an aberrant item should be adjusted appropriately.

School Location

The results showed that the location of Malaysian schools did not have a significant bearing on the derivation of the item caution indices. This indicated that the relative conceptual understanding of the subject was the same for rural and urban students. This finding also implied that the students in Malaysia were provided with the relatively same opportunities to study the subject regardless of the location of their school. This is probably promoted by the existing uniform curriculum in Malaysia.

Students' Gender

The results showed that there was no significant difference between the means of the item caution indices derived from male and female students. This finding suggested that the conceptual understanding of biology was the same for both types of students.

Students' Socio-Economic Status (SES)

There was no significant difference between the means of the item caution indices derived from students of low SES and students of middle-high (SES). This meant that the Malaysian students' SES did not have a significant influence on the students' understanding of the topics as arranged in the sequence of the items on the test.

Teachers' Working Experience

The results showed that the amount of teachers' working experience did not have a significant effect on the item caution indices. Apparently the students taught by less experienced teachers understood the concepts of the subject just as well as the students taught by more experienced teachers. This again may be influenced by the academic ability of the students, and in part, due to the uniformity of the instructional resources (same textbook) and the common curriculum followed by all the schools.

Interaction between School Setting and Students' SES

The results showed that there was no significant interaction between the student caution indices derived from students in different school settings and the students' with different SES. This implied that students from different school settings and socio-economic status had a similar conceptual understanding of the topics in the subject.

Implications

The caution index appears be to a tool that can be easily interpreted by teachers, students, parents and other interested parties. However, it appears to be influenced by the number of students taking the test and the number of items on the test. As such the results of the test administered to a single classroom may not be accurately generalized to a broader context, say for example, to the whole grade level within the school.

By virtue of arranging the items in decreasing p-values, the Sato model allows the teacher to assess the students' conceptual understanding of the topics in the subject according to the difficulty of the items. It does not provide the teacher a means to assess her students' conceptual understanding of the topics in the instructional sequence used by her when teaching the subject.

The ordering theoretic analysis of this study showed that the ordering of the items according to its prerequisite requirements was closely linked to its p-value. It was hoped that the ordering theoretic analysis would have been able to provide the teacher with alternative ways to guide her instruction according to the needs of her students. Although a conceptual map with 40 items may be mapped, it must be remembered that each item by itself does not fully represent a topic. Therefore, when using the conceptual map to aid in guiding instruction, steps must be taken to see that there are sufficient items to represent each topic tested. There would

also be a great deal of subjective judgment on the part of the teacher in deciding on the number of items for each topic and the integration of selected items under more general topics.

The type group characteristics did not significantly affect the derivation item caution indices. This maybe also influenced by the effects of uniform curriculum and the use of the same textbook for instruction.

Recommendations

There is a need for further research to be done to investigate if other factors like different instructional sequences, different textbooks, different student abilities and school attendance have an influence on the derivation of the caution indices and the probabilistic model z of the ordering theoretic analysis. It would also be useful to know the pattern of the logical hierarchy of the topics as perceived by the teachers and if it were stable over different settings. In other words, do teachers have different conceptual maps when teaching different groups of students under different settings? Are these conceptual maps also influenced by years of teaching experience and the teachers' gender?

In the Sato model, there is also the need find a parsimonious way to verify the reliability and the validity of the caution indices derived in the analysis. Perhaps a correlational analysis of the caution indices derived on parallel items on a test may provide a solution. A

conceptually sound method of resolving the arrangement of the students with the same total scores is also needed. In addition to this, there is a need to devise a more accurate way to determine an appropriate classification criterion value of the caution index. There is some evidence in this study to believe that it should not be rigidly set at .50. Research into this area should include assessing the relationship between the general quality of items on the test and the overall academic ability of the students taking the test.

Despite the work that still needs to be done in this area, the caution index appears to be a promising tool to help teachers evaluate and guide students in instruction. Combined with the ordering theoretic analysis, teachers may be able to gain further insight on their teaching practices to bring about more efficient classroom instruction. Further research into the caution indices and the ordering theory would only improve the accuracy of the inferences that may be made from these indices.

APPENDICES

APPENDIX A

THE TABLE OF SPECIFICATIONS OF THE TEST

Topic	Section I	Section II	Total
1. Size, surface, shape, support and movement	Items 1, 2	Item 26	3
2. Respiration	Items 3, 4	Item 27	3
3. Animal nutrition	Item 5	Items 28, 29*, 30	4
4. Plant nutrition	Items 6, 7	Item 31	3
5. Soil habitat and the	Items 8, 9	Item 32	3
6. Dependence of life on water	Items 10, 11	Item 33	3
7. Interdependence of living things with the environment	Items 12, 13	Item 34	3
8. Microbiology	Items 14, 15	Item 35	3
9. Interactions within a community	Items 16, 17	Item 36	3
10. Water and body fluids	Items 18, 19	Item 37	3
11. Growth, reproduction, and variation	Items 20, 21	Item 38	3
12. Detecting changes in the environment and coordination	Items 22, 23	Item 39	3
13. One organism as a habitat for another	Items 24, 25	Item 40	3
	25 items	15 items	40

Note: Item 29* can also be placed under topic 9.

APPENDIX B

THE TEST INSTRUMENT

MALAYSIA CERTIFICATE OF EDUCATION

BIOLOGY

Paper 1

One hour and fifteen minutes

Directions

1. **This test consists of 40 objective questions.**
2. **Do not open this question booklet until told to do so.**
3. **Do not make any marks on this question booklet as this booklet will be used again.**
4. **Please answer all questions.**

Section I

Directions: *Every question or incomplete statement is followed by five options. Choose* **one** *answer which you think is the best for each question and shade the appropriate space on the answer sheet.*

1. The main advantage of a curved backbone to a four-legged mammal is to

   A. spread the body weight on the backbone
   B. protect the spinal cord
   C. determine the shade of the animal
   D. protect the internal organs
   E. enable the animal to move easily

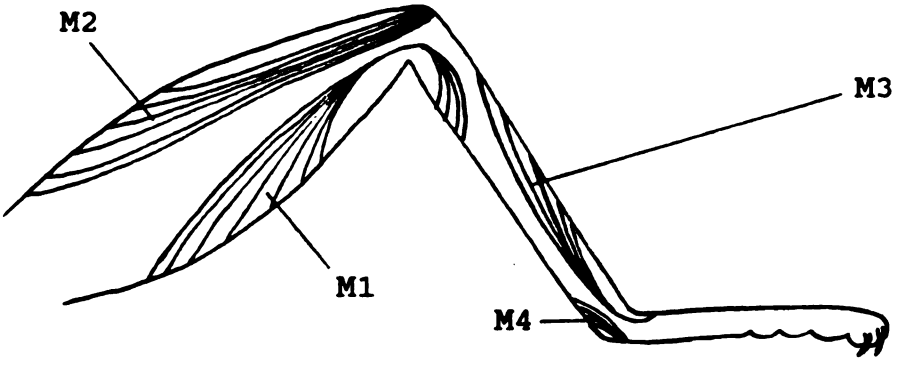


Figure 1

2. Figure 1 shows the arrangement of the muscles in the foot of an insect. In order to flex the leg, which muscles must contract?

   A. M1 and M2
   B. M1 and M3
   C. M2 and M3
   D. M2 and M4
   E. M3 and M4

3. A readily available source of energy for the contraction of muscles is obtained from

   A. oxyhaemoglobin in the blood
   B. glucose present in the muscles
   C. adenosine triphosphate in the muscles
   D. glycogen stored in the muscles
   E. fats in the body

(Turn over

4. In the fish gills, blood flows in the opposite direction to the flow of water in order to

   A. enable greater exchange of water through the gills
   B. allow efficient absorption of oxygen from the water
   C. ensure that the gills are free from dirt
   D. ease the flow of blood through the gills
   E. allow tissue respiration to occur

5. A mammal that possesses the dental formula $i\ \frac{0}{3},\ c\ \frac{0}{1},\ pm\ \frac{3}{3},\ m\ \frac{3}{3}$ most probably eats

   A. grass
   B. fish
   C. insects
   D. small mammals
   E. large mammals

6. Green plants absorb carbon dioxide and release oxygen in strong sunlight because the plants carry out

   A. photosynthesis but not respiration
   B. respiration but not photosynthesis
   C. photosynthesis faster than respiration
   D. respiration faster than photosynthesis
   E. photosynthesis and respiration at the same rate

7. A green plant that is placed in an illuminated place and supplied with carbon dioxide that contains radioactive carbon, ^{14}C. After several hours, ^{14}C is detected in the sugar in the leaves. This experiment shows that

   A. sugar is only produced by plants during the process of photosynthesis
   B. plants require radioactive carbon dioxide to produce sugar
   C. the carbon in the sugar that is produced comes from carbon dioxide
   D. carbon dioxide is necessary for the process of photosynthesis
   E. sunlight is necessary for photosynthesis

8. What happens during leaching?

   A. The humus content increases.
   B. The aeration of the soil improves.
   C. Large soil particles are broken down.
   D. Mineral salts are removed from the soil surface.
   E. The number of soil organisms increase.

(Turn over

9. After strong heating, a large sample of soil shows a small loss in weight. This means that the sample contains only a small amount of

   A. mineral salts
   B. organic substances
   C. air
   D. acidic substances
   E. alkaline substances

10. The guard cells of the leaves possess walls of differing thickness in order to

   A. prevent the entry of water through the stoma
   B. enable the opening and closing of the stoma
   C. protect the cell from injury
   D. permit uneven loss of water
   E. form a barrier towards radiation

11. The flow of air can increase the rate of transpiration because this flow

   A. carries water vapor from the leaf surface
   B. reduces the gradient of water vapor diffusion
   C. possesses a cooling effect on leaves
   D. increases the comparative air moisture in the leaves
   E. encourages the opening of stoma

12. Floating and submerged plants were found in a freshwater pond. The submerged plants gradually died because they did not obtain sufficient

   A. mineral salts
   B. space to live
   C. sunlight
   D. carbon dioxide
   E. oxygen

13. Which of the following is a quantitative experiment?

   A. The effect of light on the distribution of *Pleurococcus*
   B. The mark and release method of estimating the size of the population of woodlice
   C. The action of maltase on maltose to produce glucose
   D. Detection of the upward flow of water in a plant stem
   E. The response of a maize root tip towards light

(Turn over

14. In a microscope, when an eyepiece lens of 10x power and an objective lens of 10x are used, 80 cells are seen across the diameter of the field of vision. If the objective lens is changed to a power of 40x and the eyepiece lens remains at 10x, how many cells can be seen across the diameter of the field of vision?

A. 10
B. 20
C. 40
D. 80
E. 160

15. 1 cm^3 of river water is mixed with 99 cm^3 of sterilized distilled water. 1 cm^3 of this diluted solution is poured into 9 cm^3 nutrient agar in a sterile petri dish. If 8 colonies of bacteria are observed after incubation, how many bacteria were found in the original 1 cm^3 of river water?

A. 90
B. 792
C. 800
D. 900
E. 8000

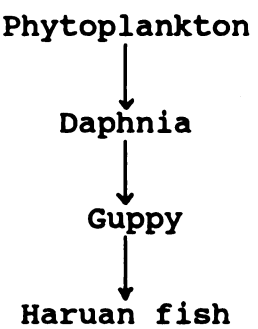


Figure 2

16. Figure 2 shows a food chain. Which of the following statements is true of this food chain?

A. There is no loss of energy from the phytoplankton to the Haruan fish.
B. The Daphnia is a type of carnivore.
C. There are more Guppy than Daphnia.
D. The Guppy is a secondary consumer.
E. All the organisms in the food chain are animals.

(Turn over

17. Among the following sequences, which one explains the ecological changes that occur on a piece of exposed land over a period of time?

A. Colonization⟶competition⟶succession⟶dominant plants
B. Succession⟶colonization⟶dominant plants⟶ competition
C. Competition⟶colonization⟶succession⟶ dominant plants
D. Dominant plants⟶colonization⟶succession⟶ competition
E. Competition⟶succession⟶dominant plants⟶ colonization

18. How does the body react when a large quantity of water is drunk all at once by a person?

A. He excretes a large quantity of dilute urine.
B. He excretes a small amount of dilute urine.
C. He excretes a normal quantity of dilute urine.
D. He excretes a normal quantity of concentrated urine.
E. He excretes a small quantity of concentrated urine.

19. Samples of blood from separate arteries were analyzed for their oxygen content. The results are shown in Table 1.

Artery	Oxygen content ($cm^3/100cm^3$ blood)
V	10.6
W	18.0
X	18.2
Y	18.5
Z	19.0

Table 1

From the results above, artery V is the

A. aorta
B. renal artery
C. intestinal artery
D. hepatic artery
E. pulmonary artery

(Turn over

20. Which one among the following parts of a plant possesses a haploid genetic constitution?

A. Cambium
B. Pollen
C. Meristem
D. Root
E. Xylem

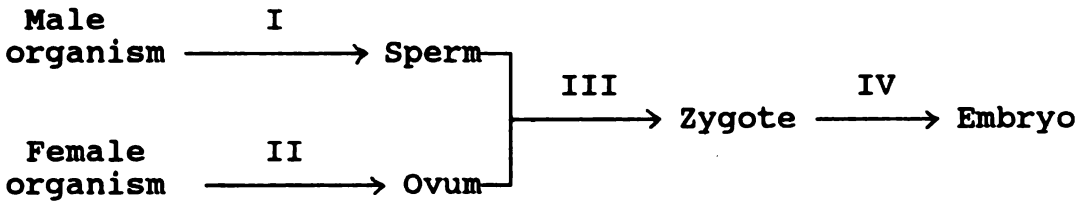


Figure 3

21. Figure 3 shows the various stages of natural reproduction. What are the processes that occur at stages I, II, III and IV?

	Stage I	Stage II	Stage III	Stage IV
A	Meiosis	Mitosis	Meiosis	Fertilization
B	Meiosis	Meiosis	Fertilization	Mitosis
C	Mitosis	Fertilization	Meiosis	Meiosis
D	Mitosis	Meiosis	Fertilization	Mitosis
E	Mitosis	Mitosis	Fertilization	Meiosis

22. A patient with diabetes (*Diabetes mellitus*) usually is treated with injections of insulin because insulin

A. stimulates the production of antibodies
B. stimulates glucose to change to glycogen
C. increases the oxidation of glucose in the intestines
D. stimulates the absorption of glucose in the intestines
E. reduces the carbohydrate metabolism

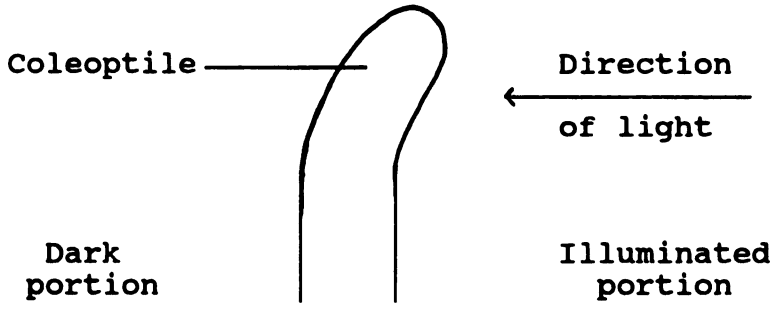


Figure 4

23. Figure 4 shows the reaction of a coleoptile towards the stimulus of light. Which of the following statements causes the coleoptile to bend?

    A. Light stimulates greater production of auxin in the coleoptile.
    B. More auxin accumulates in the illuminated portion of the coleoptile.
    C. Auxin is degenerated on the dark portion of the coleoptile.
    D. The cells on the illuminated portion of the coleoptile stops growing.
    E. The cells on the dark portion of the coleoptile elongates faster.

24. Which of the following shows an epiphytic relationship?

    A. Mucor growing on a piece of bread
    B. Mushroom growing on a wooden branch
    C. Bacteria living in the root nodules of a legume
    D. Moss growing on the bark of a tree
    E. Mistletoe growing on the branch of a tree

25. Orchard growers know that leaf bugs destroy many of their trees. Which of the following is the most suitable biological control method to eliminate this pest?

    A. Release ladybird bugs in the fruit orchards
    B. Cut off the branches that are infected
    C. Spray insecticide on the fruit trees
    D. Spray fungicide on the fruit trees
    E. Release snakes on the fruit trees

(Turn over

Section II

Directions: *For each question below,* **one** *or* **more** *of the statements are correct. Determine whether each of the statements is true or false. Then choose*

A. if I, II and III only are correct
B. if I and III only are correct
C. if II and IV only are correct
D. if IV only is correct
E. if all I, II, III and IV are correct

Directions summarized

A	B	C	D	E
I, II, III only	I, III only	II, IV only	IV only	I, II, III, IV (all four)

26. Which of the following can lower a person's body temperature when the surrounding conditions are hot?

I The relaxation of the retractor muscle of the hair
II Vasodilation of the skin's blood vessels
III An increase in the respiration rate
IV An increase in the metabolic rate

27. The outer surface of an animal's respiratory organ is thin in order to

I enable active transport of gases to occur
II increase the surface area for gaseous exchange
III reduce the formation of carbon dioxide
IV facilitate the movement of gases through this layer

28. The features of the small intestine that enable the absorption of digested food substances include

I a large surface area
II a damp surface
III possessing thin-walled villi
IV possessing layers of muscle

(Turn over

Directions summarized

A	B	C	D	E
I, II, III only	I, III only	II, IV only	IV only	I, II, III, IV (all four)

29. A number of food tests were carried out on a sample of food. The observations of the tests are shown in Table 2.

Test	Observation
Mixed with iodine solution	Yellow solution
Mixed with DCPIP solution	Blue color disappears
Heated with Millon's solution	Brick-red precipitate
Heated with Benedict'solution	Blue solution

Table 2

The observations show that the food sample contains

I protein
II reducing sugar
III vitamin C
IV starch

30. Which of the following features is(are) true for both hormones and enzymes?

I Performs specific reactions
II Controlled by temperature
III Required in small quantities for reactions
IV Produced by glands with ducts

31. The rate of photosynthesis is controlled by

I total number of leaves
II total amount of light received by the leaves
III the size of the stoma
IV the amount of oxygen in the air

(Turn over

Directions summarized

A	B	C	D	E
I, II, III only	I, III only	II, IV only	IV only	I, II, III, IV (all four)

32. The capacity of a sample of soil to retain water depends on

I the amount of humus in the soil
II the size of the soil particles
III the amount of air spaces in the soil
IV the quantity of clay in the soil

33. The importance of transpiration in plants is to

I assist the movement of water
II reduce the temperature of the plant
III assist in the absorption of mineral salts
IV prevent the wilting of leaves

34. Among the following features, which assists in the wind dispersal of fruits?

I The mesocarp layer that is hollow or fibrous
II Wing-like extensions from the growth of the pericarp
III The endocarp that is succulent and sweet
IV A tuft of hair from the remains of the pistil

35. Which of the following statements is(are) true of vaccines?

I One example of a vaccine is BCG
II Vaccines are made from pathogens that have lost their virulence
III Artificial active immunization is attained through vaccines
IV When a vaccine is injected, it will kill the pathogen

36. Which of the following occurs during succession in the area of an abandoned tin mine?

I The chemical features of the soil change
II The species of the plants change
III The species of animals change
IV The organic substances in the soil change

(Turn over

Directions summarized

A	B	C	D	E
I, II, III only	I, III only	II, IV only	IV only	I, II, III, IV (all four)

37. When a person donates blood, the doctor removes the donor's blood from a vein instead of an artery because

I the blood pressure in the vein is lower
II veins are found closer to the skin's surface
III the wall of the vein is thinner
IV the flow of blood in the vein is slower

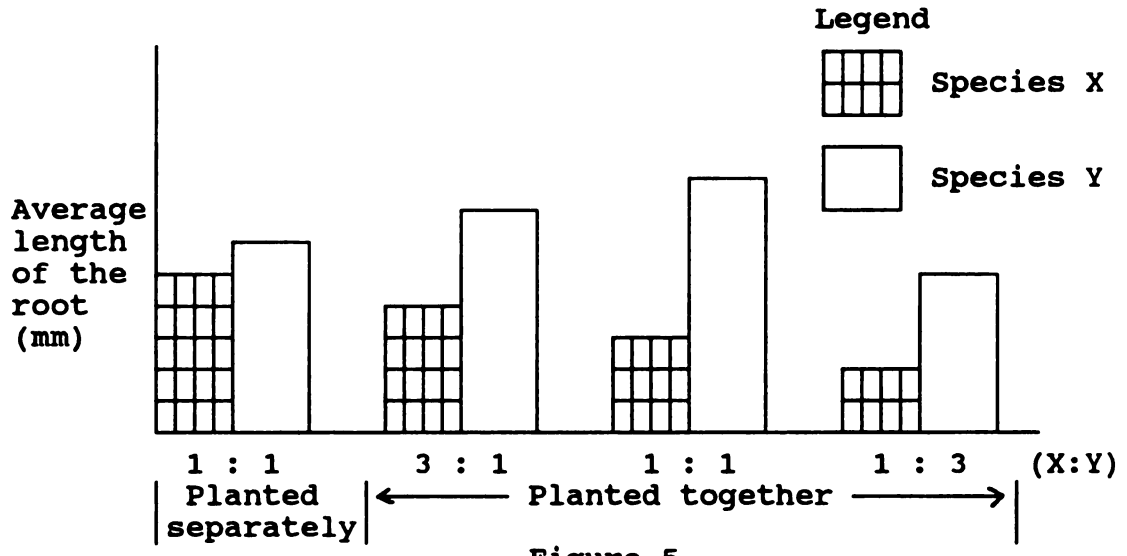


Figure 5

38. Figure 5 shows the length of the roots of two species of plants X and Y when they were planted separately and when they were planted together in different ratios, under the same conditions. What conclusion(s) can be drawn from the results that were obtained?

I The growth of the roots of species X is influenced by the number of species Y that is present.
II The growth of the roots of species Y is reduced when the number of species X increases.
III The growth of the roots of species X is reduced because of competition with the growth of the roots of species Y.
IV The growth of the roots of species Y increases when planted together with species X.

(Turn over

Directions summarized

A	B	C	D	E
I, II, III only	I, III only	II, IV only	IV only	I, II, III, IV (all four)

39. When a person sees a red flower from a distance, what changes occur in his eyes?
    I The cone cells are stimulated.
    II The focal length of the eye lenses increase.
    III The ciliary muscles relax.
    IV The eye lenses become thicker.

40. X is a green plant that carries out photosynthesis. Y and Z are two types of plant that live on the outer surface of X. Y decomposes the remains of the bark of X and this supplies mineral salts to Z. Z is able to photosynthesize. Y and Z are always found living together and receive mutual benefit from one another. Which of the following statements is(are) true regarding the living relationships and nutritional habits of the plants?
    I X is an autotroph
    II Y is a saprophyte
    III Y and Z live symbiotically
    IV Z is an epiphyte

APPENDIX C

THE TEACHER'S QUESTIONNAIRE

Please fill in the appropriate information or circle the appropriate answer. Thank you for your cooperation.

1. Name of school:.....

2. Sex: M / F

3. Academic qualifications:..... Year:.....

4. Professional qualifications:..... Year:.....

5. Number of years teaching Biology at SPM level:.....

6. What Biology classes were taught by you in 1991/1992?

   .....

7. Were the classes taught in 1992, a follow up from 1991?

   Yes / No

8. What is the name of the textbook used for instruction?

   .....

9. Did you use any supplementary texts? Yes / No

   If "Yes" what were they?.....

10. Did you teach the class(es) of 1991/92 the topics in the same sequence adopted by the textbook used? Yes / No

    If "No" what was the sequence of topics adopted?

    In 1991:.....

    .....

    In 1992:.....

    .....

11. Did you complete the syllabus? Yes / No

    If "No" what topics did you leave out?

    .....

APPENDIX D

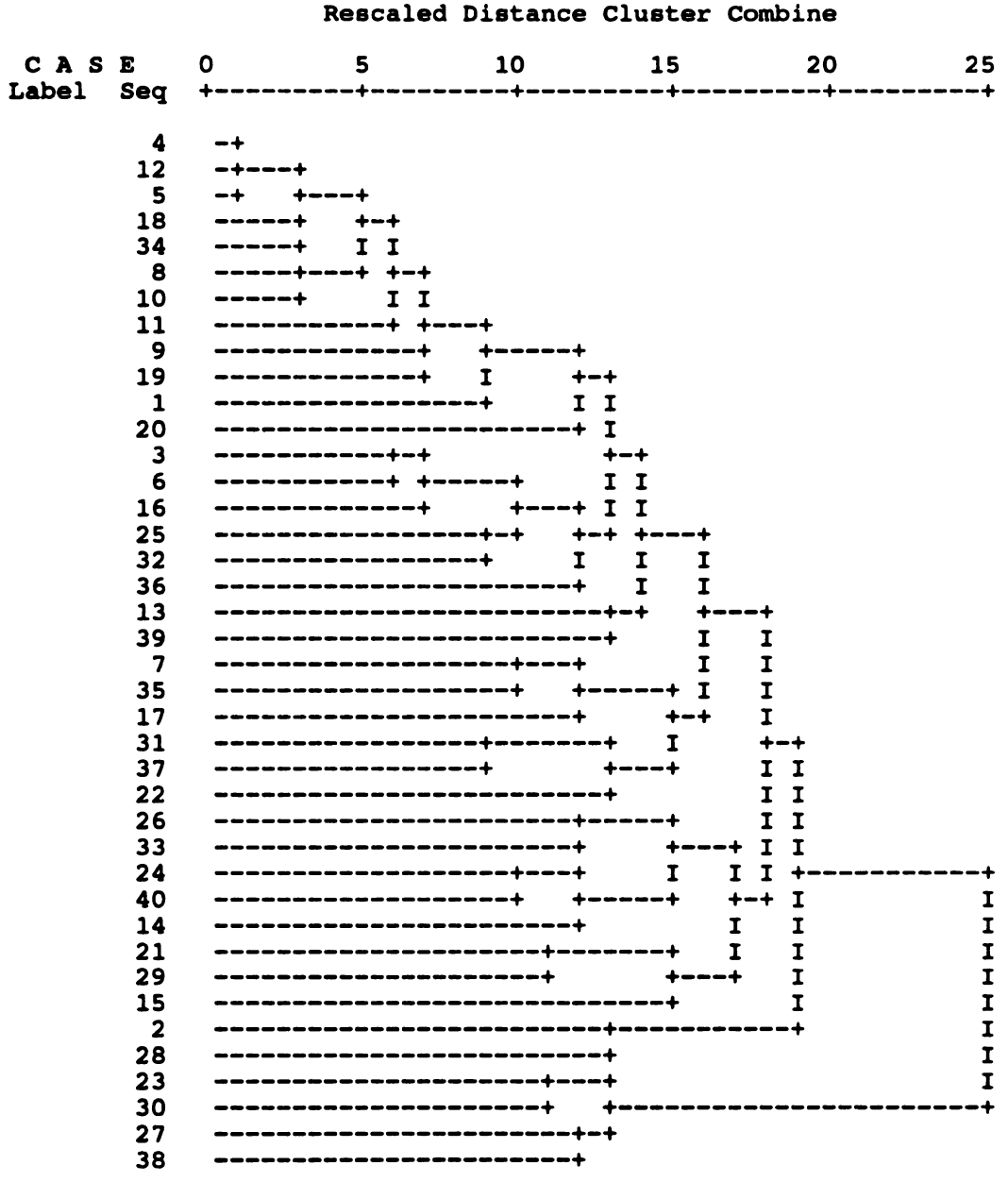
DESCRIPTIVE STATISTICS OF THE ITEMS CHOSEN FOR THE TEST INSTRUMENT

Item	Key	p	MCS	Rpbis	Year	#[a]	Topic[b]
1	A	.685	13.970	.357	1986	1	1
2	B	.635	13.244	.080	1988	1	1
3	C	.721	13.897	.360	1986	3	2
4	B	.844	13.512	.298	1989	3	2
5	A	.845	13.457	.266	1987	5	3
6	C	.595	14.635	.400	1988	4	4
7	C	.447	14.937	.435	1990	11	4
8	D	.735	14.096	.456	1987	9	5
9	B	.727	13.952	.388	1988	9	5
10	B	.774	13.986	.455	1988	6	6
11	A	.660	14.518	.528	1986	11	6
12	C	.857	13.461	.282	1987	17	7
13	B	.594	13.693	.209	1990	12	7
14	B	.411	14.353	.282	1989	14	8
15	C	.619	13.630	.200	1990	14	8
16	D	.810	13.699	.359	1990	17	9
17	A	.704	13.879	.338	1989	13	9
18	A	.763	13.675	.301	1986	18	10
19	E	.675	14.305	.470	1989	18	10
20	B	.659	14.422	.493	1987	21	11
21	B	.669	14.283	.455	1989	20	11
22	B	.701	14.011	.386	1987	24	12
23	E	.439	15.223	.491	1990	22	12
24	D	.518	14.837	.476	1988	23	13
25	A	.887	13.422	.294	1989	25	13
26	A	.501	14.556	.390	1988	26	1
27	D	.325	14.872	.325	1988	28	2
28	A	.588	13.782	.233	1988	29	3
29	B	.712	14.303	.511	1990	29	3&9
30	A	.268	13.728	.109	1990	39	3
31	A	.642	14.050	.351	1988	31	4
32	E	.773	13.378	.173	1986	31	5
33	A	.431	15.012	.438	1989	33	6
34	C	.872	13.504	.329	1986	34	7
35	A	.647	14.448	.490	1990	34	8
36	E	.609	14.116	.348	1987	34	9
37	E	.388	14.559	.309	1989	37	10
38	E	.388	14.559	.309	1989	36	11
39	A	.663	14.097	.384	1989	38	12
40	E	.631	14.146	.374	1989	39	13

[a] This refers to the question number as it appeared in the national exam of that year.
[b] The topic number refers to the topics as listed in the table of specifications (see Appendix A).

APPENDIX E

DENDROGRAM OF THE AGGLOMERATIVE HIERARCHICAL CLUSTER ANALYSIS OF THE TEST ITEMS USING THE COMPLETE LINKAGE METHOD



APPENDIX F

THE ITEM CAUTION INDICES OF THE 40 ITEMS OF THE TEST INSTRUMENT AS DERIVED UNDER VARIOUS SAMPLES

ITEM NUMBER	WHOLE SAMPLE	TYPE OF SAMPLE SCHOOL SETTING URBAN	RURAL	STUDENT SEX MALE	FEMALE	STUDENT SES LOW	M/H
1	.74	.77	.76	.85	.62	.77	.74
2	1.00	.91	1.03	1.05	.99	1.01	.98
3	.55	.57	.55	.59	.50	.50	.58
4	.47	.47	.52	.67	.41	.50	.45
5	.59	.53	.68	.41	.75	.66	.52
6	.37	.52	.32	.47	.30	.29	.50
7	.44	.46	.50	.46	.42	.39	.50
8	.32	.26	.41	.28	.37	.30	.39
9	.45	.44	.41	.42	.49	.39	.50
10	.45	.57	.35	.39	.49	.31	.67
11	.45	.38	.55	.48	.39	.36	.58
12	.57	.55	.65	.67	.52	.57	.61
13	.74	.67	.77	.80	.67	.81	.66
14	.54	.65	.46	.62	.60	.49	.59
15	.88	.69	.99	.79	.94	.99	.73
16	.48	.52	.48	.47	.48	.40	.57
17	.66	.61	.64	.63	.68	.60	.69
18	.56	.68	.55	.73	.36	.48	.64
19	.41	.53	.36	.39	.40	.39	.47
20	.59	.57	.61	.57	.61	.57	.63
21	.66	.67	.65	.61	.68	.59	.76
22	.69	.66	.77	.62	.76	.76	.64
23	.65	.70	.57	.76	.52	.48	.83
24	.50	.51	.48	.61	.36	.52	.47
25	.69	.69	.45	.79	.63	.74	.60
26	.60	.64	.64	.77	.46	.63	.59
27	.65	.67	.58	.72	.57	.55	.76
28	.79	.79	.84	.69	.90	.80	.78
29	.48	.56	.41	.46	.49	.38	.59
30	.91	.82	.99	.87	.94	.98	.81
31	.53	.59	.47	.58	.49	.48	.55
32	.68	.65	.71	.83	.53	.61	.71
33	.49	.51	.52	.47	.54	.52	.47
34	.52	.55	.55	.47	.56	.48	.57
35	.41	.49	.33	.45	.38	.42	.38
36	.65	.76	.53	.63	.66	.59	.77
37	.62	.74	.55	.65	.60	.52	.70
38	.71	.66	.79	.61	.88	.70	.74
39	.60	.61	.53	.58	.64	.63	.57
40	.54	.50	.57	.53	.50	.61	.45
n	354	192	162	198	156	173	180

APPENDIX F (CONT'D).

THE ITEM CAUTION INDICES OF THE 40 ITEMS OF THE TEST INSTRUMENT AS DERIVED UNDER VARIOUS SAMPLES

ITEM NUMBER	WHOLE SAMPLE	TYPE OF SAMPLE TEACHERS' EXPERIENCE 0-5 YRS	0-10 YRS	>10 YRS	TEST FORMAT[a]
1	.74	.62	.69	.79	.66
2	1.00	1.11	1.03	.95	1.01
3	.55	.48	.46	.64	.51
4	.47	.32	.43	.52	.51
5	.59	.83	.66	.57	.65
6	.37	.54	.40	.34	.41
7	.44	.54	.51	.36	.42
8	.32	.19	.29	.34	.33
9	.45	.45	.44	.46	.45
10	.45	.77	.65	.28	.40
11	.45	.37	.41	.48	.41
12	.57	.65	.60	.55	.53
13	.74	.42	.61	.87	.68
14	.54	.59	.50	.58	.49
15	.88	.70	.90	.86	.80
16	.48	.58	.53	.44	.45
17	.66	.51	.59	.72	.58
18	.56	.63	.60	.52	.59
19	.41	.49	.42	.40	.40
20	.59	.74	.63	.54	.55
21	.66	.69	.70	.63	.64
22	.69	.56	.61	.75	.62
23	.65	.49	.57	.74	.60
24	.50	.54	.42	.58	.49
25	.69	.63	.64	.73	.69
26	.60	.72	.68	.53	.53
27	.65	.71	.60	.71	.59
28	.79	.84	.73	.84	.68
29	.48	.52	.49	.46	.43
30	.91	.70	.73	1.15	.78
31	.53	.61	.48	.57	.47
32	.68	.69	.72	.64	.56
33	.49	.56	.48	.51	.42
34	.52	.63	.56	.46	.44
35	.41	.45	.40	.42	.35
36	.65	.66	.61	.69	.51
37	.62	.59	.49	.75	.52
38	.71	.76	.73	.69	.64
39	.60	.46	.45	.76	.54
40	.54	.53	.54	.53	.47
n	354	198	156	173	354

[a]These are the combined values for the first format (Section I) and second format (Section II).

BIBLIOGRAPHY

BIBLIOGRAPHY

Airasian, P. W. (1971). A method for validating sequential instructional hierarchies. Educational Technology, 11(12), 54-56.

Airasian, P. W., & Bart, W. M. (1973). Ordering theory: A new and useful measurement model. Educational Technology, 23(5), 56-60.

Airasian, P. W., & Bart, W. M. (1975). Validating a priori instructional hierarchies. Journal of Educational Measurement, 12(3), 163-173.

Airasian, P.W., Bart, W. M., & Greaney, B. J. (1975). The analysis of a proportional logic game by ordering theory. Child Study Journal, 5(1), 13-24.

Airasian, P. W., & Madaus, G. F. (1983). Linking testing and instruction: Policy issues. Journal of Educational Measurement, 20(2), 103-118.

Baker, E. L., & Herman, J. L. (1983). Task structure design: Beyond linkage. Journal of Educational Measurement, 20, 149-164.

Bart, W. M. (1974). Test validity and reliability from an ordering-theoretic framework. Educational Technology, 24(1), 62-63.

Bart, W. M. (1978). An empirical inquiry into the relationship between test factor structure and test hierarchical structure. Applied Psychological Measurement, 2(3), 331-335.

Bart, W. M., & Airasian, P. W. (1974). Determination of the ordering among seven Piagetian tasks by an ordering-theoretic method. Journal of Educational Psychology, 66, 277-284.

Bart, W. M., Frey, S., & Baxter, J. (1979). Generalizability of the ordering among five formal reasoning tasks by an ordering-theoretic method. Child Study Journal, 9, 251-259.

Bart, W. M., & Mertens, D. M. (1979). The hierarchical structure of formal operational tasks. Applied Psychological Measurement, 3(3), 343-350.

Bart, W. M., & Krus, D. J. (1973). An ordering-theoretic method to determine hierarchies among items. Educational and Psychological Measurement, 33, 291-300.

Bejar, I. (1984). Educational diagnostic assessment. Journal of Educational Measurement, 21(2), 175-189.

Birenbaum, M., & Shaw, D. J. (1985). Task specification chart: A key to a better understanding of test results. Journal of Educational Measurement, 22(3), 219-230.

Birenbaum, M., & Tatsuoka, K. K. (1982). On dimensionality of achievement test data. Journal of Educational Measurement, 19(4), 259-266.

Blixt, S. L., & Dinero, T. E. (1985). An initial look at the validity of diagnoses based on Sato's caution index. Educational and Psychological Measurement, 45, 293-299.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. Applied Psychological Measurement, 12, 261-280.

Brown, J. S., & Burton, R. R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. Cognitive Science, 2, 155-192.

Clark, C. M., & Peterson, P. L. (1986). Teachers' thought processes. In M. C. Wittrock (Ed.), Third handbook of research on teaching (pp. 225-296). New York: Macmillan.

Donlon, T. F., & Fischer, F. E. (1968). An index of an individual's agreement with group-determined item difficulties. Educational and Psychological Measurement, 28, 105-113.

Drasgow, F. (1978). Statistical indices of the appropriateness of aptitude test scores. Unpublished doctoral dissertation, University of Illnois, Urbana-Champaign.

Druva, C. A. (1985, April). A composite of order analysis procedures. A paper presented at the American Educational Research Association National Convention at Chicago, Illinois. (Eric Document Reproduction Service No. ED 275 764)

Ebel, R. E. (1972). Essentials of educational measurement (3rd ed.). New Jersey: Prentice-Hall, Inc.

Egan, O., & Archer, P. (1985). The accuracy of teachers' ratings of ability: A regression model. American Educational Research Journal, 22, 25-34.

Floden, R. E., Porter, A. C., Schmidt, W. H., & Freeman, D. J. (1980). Don't they all measure the same thing: Consequences of standardized test selection (pp.109-120). In E. L. Baker, & E. S. Quellmalz (Eds.), Educational testing and evaluation. Beverly Hills, CA: SAGE Publications.

Fuchs, L. S., & Fuchs, D. (1986). Linking assessment to instructional intervention: An overview. School Psychology Review, 15(3), 318-323.

Fujita, H., Satoh, T., & Nagaoka, K. (1977). Graphical analysis of test scores using an S-P table. Educational Technological Research, 1, 21-31.

Gagne', R. M. (1985). The conditions of learning and theory of instruction (4th ed.). New York: Holt, Rinehart & Winston.

Green, S. B. (1983). Identifiability of spurious factors using linear factor analysis with binary items. Applied Psychological Measurement, 7(2), 139-147.

Guttman, L. A. (1950). A basis for scalogram analysis. In S.A. Stouffer et al. (Eds.), Studies in social psychology in World War II: Measurement and prediction (Vol. 4, pp. 60-90). Princeton, NJ: Princeton University Press.

Harnisch, D. L. (1983). Item response patterns: Applications for educational practice. Journal of Educational Measurement, 20(2), 191-206.

Harnisch, D. L., & Linn, R. L. (1981). An analysis of item response patterns: Questionable test data and dissimilar curriculum practices. Journal of Educational Measurement, 18(3), 133-146.

Harnisch, D. L.., & Romy, N. (1985). User's guide for the Student Problem Package (SPP) on the IBM-PC. University of Illinois at Urbana- Champaign, Office of Educational Testing, Research and Service, Champaign, Illinois.

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. Review of Educational Research, 59(3), 297-313.

Jaeger, R. M. (1988). Use and effect of caution indices in detecting aberrant patterns of standard-setting judgments. Applied Measurement in Education, 1(1), 17-31.

Kane, M. T., & Brennan, R. L. (1980). Agreement coefficients as indices of dependability for domain-referenced tests. Applied Psychological Measurement, 4(1), 105-126.

Krus, D. J. (1974). A computer program for deterministic and probabilistic models of order analysis. Educational and Psychological Measurement, 34, 677-683.

Krus, D. J. (1975). Order analysis of binary data matrices. Los Angeles: Theta Press.

Krus, D. J. (1977). Order analysis: An inferential model of dimensional analysis and scaling. Educational and Psychological Measurement, 37, 587-601.

Krus, D. J. (1978). Logical basis of dimensionality. Applied Psychological Measurement, 2(3), 321-329.

Krus, D. J., & Bart, W. M. (1974). An ordering theoretic method of multidimensional scaling of items. Educational and Psychological Measurement, 34, 525-535.

Krus, D. J., Bart, W. M., & Airasian, P. W. (1975). Ordering theory and methods. Theta Press.

Krus, D. J., & Krus, P. H. (1980). Dimensionality of Hierarchical and proximal data structures. Applied Psychological Measurement, 4(3), 313-321.

Krus, D. J., & Weiss, D. J. (1976). Empirical comparison of factor and order analysis on prestructured and random data. Multivariate Behavioral Research, 11, 95-104.

Leinhardt, G., & Seewald, A. M. (1981). Student-level observation of beginning reading. Journal of Educational Measurement, 18(3), 171-178.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple choice test scores. Journal of Educational Statistics, 4, 269-290.

Linn, R. L. (1983). Testing and instruction: Links and distinctions. Journal of Educational Measurement, 20, 180-189.

Linn, R. L. (1990). Essentials of student assessment: From accountability to instructional aid. Teachers College Record, 91(3), 422-436.

McArthur, D. L. (Ed.). (1987). Alternative approaches to the assessment of achievement. Boston: Kluwer Academic Publishers.

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

Mehrens, W. A., & Lehmann, I. J. (1987). Using standardized tests in education, (4th ed.). New York: Longman Inc.

Mehrens, W. A., & Lehmann, I. J. (1991). Measurement and evaluation in education and psychology, (4th ed.). New York: Holt, Rinehart, and Winston.

Nunnally, J. C. (1978). Psychometric theory (2nd ed.). New York: McGraw Hill.

Piazza, N. J., & Wise, S. L. (1988). An order-theoretic analysis of Jellinek's disease model of alcoholism. The International Journal of Addictions, 23(4), 387-397.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics, 4(3), 207-230.

Reckase, M. D. (1985). The difficulty of test items that measure more than more ability. Applied Psychological Measurement, 9, 401-412.

Resnick, L. B. (1976). task analysis in instructional design: Some cases from mathematics. In D, Klahr (Ed.), Cognition and Instruction. Hilldsale, NJ: LEA.

Reynolds, T. J. (1981). ERGO: A new approach to multidimensional item analysis. Educational and Psychological Measurement, 41, 643-659.

Sato, T. (1980). The S-P chart and the caution index. C & C Systems Research Laboratories, Nippon Electric Co., Ltd.

Sato, T. (1990). The S-P chart analysis. In D. L. Harnisch & M. L. Connell (Translation Eds.), An introduction to educational information technology (pp. 159-175). Japan: NEC Technical College.

Sato, T., & Kurata, M. (1977). Basic S-P score table characteristics. NEC Research and Development, 47, 64-71.

Schmidt, W. H. (1983). Content biases in achievement tests. Journal of Educational Measurement, 20, 165-177.

Shishido, J. A., Ayabe, H. I., & Heim, M. (1986). Relationship between caution indices and student demographic data in a Japanese language placement examination situation. In Kim Chul-hwan and Lee Wha-kuk (Eds.), Higher education and the Asia-Pacific century. Proceedings of the '88 PRAHE Seoul Conference. (ERIC Document Reproduction Service No. ED 311 753)

Switzer, D. M., & Connell, M. L. (1990). Practical applications of student response analysis. Educational Measurement: Issues and Practice, 9(2),15-18.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. Psychometrika, 49(1), 95-110.

Tatsuoka, M. M. (1978). Recent psychometric developments in Japan: Engineers tackle educational measurement problems. Paper presented at the ONR Contractors Meeting on Individualized Measurement, Columbia, MO.

Tatsuoka, K. K., & Baillie, R. (1982). SIGNBUG: An error diagnostic computer program for signed-number arithmetic on the PLATO system. Urbana-Champaign, Illinois: University of Illinois Computer-based Research Laboratory.

Tatsuoka, K. K., & Linn, R. L. (1983). Indices for detecting unusual patterns: Links between two general approaches and potential applications. Applied Psychological Measurement, 7(1), 81-96.

Tatsuoka, K. K., & Tatsuoka, M. M. (1982). Detection of aberrant response patterns and their effect on dimensionality. Journal of Educational Statistics, 7(3) 215-231.

Tatsuoka, K. K., & Tatsuoka, M. M. (1983). Spotting erroneous rules of operation by the individual consistency index. Journal of Educational Measurement, 20(3), 221-230.

Tomsic, M. L. (1987, April). The effect of poor fitting items on the distributions of extended caution indices. A paper presented at the annual meeting of the AERA at Washington, DC.

van der Flier, H. (1982). Deviant response patterns and comparability of test scores. Journal of Cross-Cultural Psychology, 13(3), 267-298.

Wise, S. L. (1981). A modified order-analysis procedure for determining unidimensional item sets. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

Wise, S. L. (1983). Comparisons of order analysis and factor analysis in assesing the dimensionality of binary data. Applied Psychological Measurement, 7(3), 311-321.

Wise, S. L. (1986). The use of ordering theory in the measurement of student development. Journal of College Personnel, 27(5), 442-447.

Wise, S. L., & Tatsuoka, M. M. (1986). Assessing the dimensionality of dichotomous data using modified order analysis. Educational and Psychological Measurement, 46, pp.295-301.

Wright, B. D. (1979). Solving measurement problems with the Rasch model. Journal of Educational Measurement, 14, 97-116.

Zimowski, M. F., & Bock, R. D. (1987). Full-information item factor analysis of test items from the ASVAB CAT pool. Chicago: National Opinion Research Center.