COMPARING 12 FINITE STATE MODELS OF EXAMINEE PERFORMANCE ON MULTIPLE-CHOICE TESTS

by

Than Than Zin

Dissertation submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Educational Research and Evaluation

APPROVED:

James C. Impara, Chairman

Robert B. Frary

Miguel A. García-Pérez

Robert S. Schulman

Larry J. Weber

Lee M. Wolfle

April, 1992

Blacksburg, Virginia

COMPARING 12 FINITE STATE MODELS OF EXAMINEE PERFORMANCE ON MULTIPLE-CHOICE TESTS

by

Than Than Zin

Committee Chairman: James C. Impara

Educaional Research and Evaluation

(ABSTRACT)

Finite state test theory models the response behavior of an examinee and establishes the relationship between the ability of the examinee and the observed responses on a multiple-choice test. In finite state modeling, various assumptions about item characteristics and the examinees' response strategies are made to estimate an examinee ability, and willingness to guess.

Twelve sets of plausible assumptions about identifiability of distractors and examinee guessing strategies were adopted and the corresponding finite state models were actualized. Three consequences of the adoption of the 12 sets of assumptions were investigated: 1) the extent to which the resulting ability estimates rank ordered the examinees similarly, 2) variation in the magnitude of ability estimates and the estimated willingness to guess across the 12 models, and 3) the extent to which conclusions about examinees subgroups would differ according to the model employed. Also, conventional number-right scores were compared with the finite state scores with respect to the three outcomes just listed.

All scoring methods rank ordered the examinees essentially the same. The magnitude of the finite state scores varied considerably across models mainly due to differing assumptions about the identifiability of distractors. Differing assumptions about examinee guessing strategy had surprisingly little effect on the magnitude of the ability estimates, though estimates of willingness to guess varied consistently according to the assumed strategies. Conclusions about group differences also varied across the models as a result of differing assumptions about both item characteristics and examinee guessing strategies.

Acknowledgements

I hereby express my sincere gratitude to my advisor, Dr. James C. Impara, for his support and guidance throughout my study program. I am indebted to Dr. Miguel A. García-Pérez and Dr. Robert B. Frary for their assistance and many helpful suggestions during the writing of this dissertation.

I also wish to thank to other members of my committee: Dr. Robert S. Schulman, Dr. Larry J. Weber, and Dr. Lee M. Wolfle, for their suggestions and their valuable time. The data for the study were provided by the Virginia Department of Education. I am grateful to Dr. David R. Mott for his help in this regard.

Table of Contents

List of Tables

List of figures

CHAPTER I

Introduction

Most measurement experts and testing organizations prefer to use the multiple choice format for standardized cognitive tests in education and other areas. It is also one of the most common and favored test item types for classroom teachers (Marso, 1985). Characteristics of multiple-choice tests that have led to this acceptance include capacity for broad content sampling, high reliability, ease of administration and scoring, and usefulness in testing varied content. Contrary to popular belief, multiple-choice tests are not limited to testing recognition and rote memorization but may also require types of higher level thinking (Haladyna & Downing, 1989).

However, the scoring of multiple-choice tests has been problematical. In any kind of measuring process, an individual is assigned a number as a quantitative description based on his/her responses to a test. This number represents the degree to which the examinee manifests some property that the test is intended to measure. Therefore, how to give meaning to the specific number used as a test score is a very important matter (Ghiselli, Campbell, & Zedack, 1981, chap. 3).

Number-right scoring of a multiple-choice test simply ascertains whether the right answer was supplied for each item; the total number of right answers represents the examinee's score. A total score is interpreted as the degree of an examinee's knowledge or skill. However, number-right scores from multiple-choice tests are usually not directly relatable to some degree of achievement, except in the case of highly homogeneous tests. Number-right scores would be misleading if they were interpreted in the criterion-referenced sense, i.e., as percentages of some body of a

subject matter known (García-Pérez and Frary, 1989). Even when norming data are available, the information that an examinee's number-right score is 30 tells us little or nothing about the individual in terms of what he/she knows or can do. It provides only ranking information.

Number-right scoring has also been criticized for one of its underlying assumptions about examinee's response behavior. It implicitly assumes that every examinee answers all items and guesses among two or more options when the answer is unknown. Therefore, examinees should be encouraged to answer all questions under conventional number-right scoring, even if their responses for unknown items are random guesses. However, the directions provided on some tests which are scored number-right do not encourage guessing to the extent they should. The directions for many tests are ambiguous about whether examinees should guess or not when they do not know the correct answer with assurance. In fact, the test directions should specify the optimal guessing strategy clearly. Fisher (1988) reported that students' understanding of test directions is a significant factor in determining the extent to which they guess.

When test directions are ambiguous about guessing strategy and omissions occur in the responses, number-right scoring may penalize examinees who fail to guess not realizing that providing an answer to every question under number-right scoring is beneficial. Moreover, if this omissive behavior is related to some

characteristics of examinees, test scores may have differential validity both for individuals and groups of examinees.

Number-right scoring also causes irregularities in the interpretation of test scores because it does not take into account the possibility of guessing, partial knowledge and misinformation. A scoring model, based on the assumption that an examinee who knows the answer gives it and otherwise either omits the item or guesses at random is not realistic and rarely accurate (Lord & Novick, 1968, chap. 14). This model is the basis for the conventional "correction for guessing", which has been shown to be unsatisfactory when measuring different types of ability (Bliss, 1980; Cross & Frary, 1977). Examinees who have partial knowledge about an item do not guess at random among all of the options nor do examinees with misinformation about an item. For over 50 years, the substantial weight of evidence has shown that examinees are able to make use of partial knowledge and productive guessing strategies when responding to multiple-choice items (Hutchinson, 1985).

All widely used scoring techniques fail to extract potentially available information from multiple-choice test responses. Thus, introduction of more refined scoring methods, which attempt to recover additional residual information is desirable. Techniques such as response weighting (Claudy, 1978), subset selection (Jaradat & Sawaged, 1986), answer until correct (Wilcox & Wilcox, 1988), and admissible probability measurement scoring (Bruno, 1986) have been developed for this purpose and investigated. To varying extents, they have been shown to provide more or better

information about examinees than conventional number-right scoring procedures. However, their effectiveness has not been uniform and many difficulties arise in applying them (Frary, 1989). Moreover, these alternative scoring models fail to provide a comprehensive picture of response behaviors because the assumptions made in each scoring model are specific for the particular response mode for which the model is designed (García-Pérez & Frary, 1991a).

García-Pérez (1987) proposed a finite state theory of performance on multiple-choice tests using a more generalized modeling approach that provides an alternative scoring method. Compared to the conventional scoring schemes, scores derived from the finite state models have two major advantages. First, finite state theory parsimoniously incorporates assumptions appropriate to the particular mode in which the test is administered. These assumptions concern the subject's item response behavior when responding to a multiple-choice item rather than merely accounting for the actual response. Second, the theory yields ability estimates on a single metric, the same regardless of the mode of test administration or the assumptions concerning examinee behavior. Finite state scoring also considers an examinee's guessing strategies in the ability estimation procedure and provides a quantitative estimate of an examinee's willingness to guess based on the response pattern.

Item response theory (IRT) also estimates examinees' ability from the responses and provides ability estimates on an interval scale. However, the IRT approach requires complex operations of mathematics with large and appropriate

samples of examinees. During the past two decades, the application of IRT has become practical with the developments of computers and sophisticated software (Suen, 1990, chap. 7). At present, the IRT approach seems to be progressing in many promising directions. Nevertheless, the absence of inherently meaningful units of measurement in the ability scale of IRT makes interpretation difficult and the reporting of estimated ability to examinees an uncertain enterprise (Loyd, 1988). Compared to the IRT approach, finite state scoring is relatively uncomplicated and does not require advanced mathematical techniques. Moreover, finite state scoring provides a sound interpretation of test scores that is critical to determining the educational implications of the scores.

Another approach to evaluating and reporting test scores from a multiple-choice test is referred to as domain-referenced testing, in which test scores are interpreted as estimates of performance on a universe of similar items (Worthen & Sanders, 1987, chap. 18). Though a domain-referenced test may be scored number-right, the score of a specific examinee can be interpreted without reference to the scores of any other examinees. However, this approach estimates only the proportion of items in the universe of items that the examinee would answer correctly and does not reveal the actual knowledge of the examinee in any well-defined sense. In contrast, finite state theory yields scores that are directly interpretable as the amount of knowledge held by the examinee. Moreover, the definition of knowledge in finite state theory is such that the educational implications of the scores may be evaluated

unambiguously. The details of finite state theory are presented more comprehensively in Chapter 2.

The simulation study by García-Pérez and Frary (1989) was highly supportive of the use of finite state scores. The finite state approach allows us to make different assumptions about item characteristics and guessing strategies, thus describing different test-taking behaviors. When examinees take a multiple-choice test, there are several behaviors they might adopt regardless of the number of choices and type of test instructions. These behaviors may result from personal idiosyncracies, the influence of test instructions, and the examinee's knowledge of optimal guessing strategy (García-Pérez & Frary, 1989).

In this study, 12 finite state models were developed based on 12 sets of plausible assumptions about the identifiability of distractors and variations in guessing strategy. Although goodness of fit between the models incorporating different assumptions and data can be tested if the number of independent response categories exceeds the number of parameters in the models, the data in this study do not provide this situation. Therefore, throughout the study, each model was applied to the whole sample of examinees as if every examinee had followed the uniform behavior assumed in the model. Two examinee parameters, ability and willingness to guess, were estimated from each model and compared across the models to investigate how the variations of assumptions affected the estimates.

Three consequences of the adoption of 12 sets of assumptions were investigated: (1) the extent to which the resulting estimates rank ordered the examinees similarly, (2) variation in the magnitude of the ability estimates and the estimates of willingness to guess across the 12 models, and (3) the extent to which conclusions about examinee subgroups would differ according to the model employed. Also, conventional number-right scores were compared with the finite state scores with respect to the three outcomes listed above.

In particular, the following research questions were investigated regarding ability estimates.

1. To what extent does the ranking of examinees based on number-right scores tend to agree with those based on finite state scores?
2. How similar are the rankings of examinees based on finite state scores across the 12 models?
3. Do the finite state scores from different models provide similar estimates of the examinees' abilities?
4. Do conclusions about group differences revealed by finite state scores differ across the 12 models?

To explore the characteristics of estimated willingness to guess, the following research questions were investigated.

5. How similar are the estimates of examinees' willingness to guess across the 12 finite state models?

6. What kind of relationships hold between the estimates of examinees' willingness to guess and other possible indicators of omissiveness?

7. How do the estimates of willingness to guess relate to estimated ability (finite state scores) and number-right scores?

8. Do conclusions about group differences in estimated willingness to guess differ across the models?

CHAPTER II

Finite State Modeling and Test-taking Behaviors

1. Finite State Test Theory

Finite state test theory models response behavior of an examinee and establishes the relationship between the ability of the examinee and the observed responses on a multiple choice test by tracking an examinee's test-taking behavior. In the finite state approach, examinees' knowledge at the option level instead of knowledge at the item level is considered to account for partial knowledge. Partial knowledge is defined as the ability to identify some, but not all, of the distractors thus restricting guessing to a proper subset of options that includes the correct option.

The two basic parameters of finite state theory are the examinee's level of knowledge or ability and propensity for guessing. The theory defines the level of knowledge of an examinee as the proportion of statements about a subject matter whose truth value the examinee knows. The examinee's level of knowledge is represented by the symbol λ ($0 \leq \lambda \leq 1$). On the test, a statement is a sentence resulting from adding the item stem to one of its available options. Therefore, λ is the probability that the examinee will be able to classify (correctly) a randomly chosen option as a true or false completion of the item stem.

The second characteristic of the subjects is willingness to guess at random among unclassified options. The parameter γ ($0 \leq \gamma \leq 1$) is the probability of guessing at random among the unclassified options (in the absence of assured knowledge). It is intended to account for the circumstances that may lead an

examinee to guess when he/she does not know the answer. This parameter is included in the theory to invoke the fact that individual differences exist regarding examinees' actual willingness to guess rather than omit an item.

Finite state theory assumes that an examinee (in effect if not actually) makes independent attempts to classify every option of an item as true or false, and the success or failure of these attempts determines the examinee's state of knowledge about the item. Assuming that an examinee is not misinformed, i.e., having excluded the (single) correct option from consideration, these states range from total ignorance through several degrees of partial knowledge to total knowledge. The theory acknowledges that if the examinee does not have enough knowledge to select the correct answer with assurance, he/she may guess at random among the unclassified options.

Finite state theory allows us to make assumptions about item characteristics and the examinee's response strategy to provide equations for the probability of every response category that occurs under any format and administration mode of a multiple-choice test. Different finite state models can be established by changing the assumptions about item characteristics and examinee response behaviors. Other variables to be considered in finite state modeling are the response strategy followed by the examinees, the format of administration of the test, and potentially other item characteristics such as the number of options per item, and the presence of noncontent options (e.g., "none of the above").

Although an approach to accounting for misinformation was discussed in García-Pérez and Frary (1991b), it is not invoked in this study. Speed, the examinee characteristic that dictates the extent to which items at the end of the test are not reached within the time limit, was also discussed by García-Pérez and Frary (1991b), but it is not considered in this study. However, the theory as employed in this study incorporates total knowledge, partial knowledge, total ignorance, and guessing strategy. Further detailed explanations about theory, definitions and justification for some assumptions are given in García-Pérez (1985, 1987, 1989, 1990) and García-Pérez and Frary (1989, 1991a, 1991b).

2. Development of finite state models

García-Pérez and Frary (1989) showed how any assumptions representing a characteristic of the test item, the testing situation, and guessing strategy can be incorporated into a model that provides the basis for estimating each examinee's λ. The preliminary assumptions for the development of finite state models in this study are as follows:

1) local independence across items.

2) independence of options. This means that options within an item must be independently classifiable by examinees as if they were independent true false items.

3) conventional administration of the test, that is, asking examinees to mark the single option believed to be correct for each item.

4) all items have four response choices, one correct and three incorrect.

5) test instructions are incomplete regarding guessing (see Chapter 3, p. 24). The instructions do not specifically urge the examinees to answer every item, even though the test is to be scored number-right, and are ambiguous about what the examinee should do about skipped items when time is up.

Insert Figure 1 about here

Theoretical development of finite state models can be explained with a tree diagram (Figure 1). The tree diagram shows how λ and γ interact to determine whether an examinee answers correctly, incorrectly, or omits an item under the assumptions one to five above. This diagram represents all possible sequences of events when an examinee confronts a four-choice item under the instruction to select the one believed to be correct. According to the theory, for a test with four (N) options per item test, there are five (N+1) possible states of knowledge, ranging from total ignorance to total knowledge. There are three (N-1) states of partial knowledge corresponding to having or not having the knowledge needed to classify 1, 2, 3 (1, 2, ..., N-1) options regardless of whether the single correct answer is among them. Thus, five states of knowledge arise for a four-option item. These are represented in the main body of the diagram by the five rows from link one to four. The topmost path represents total knowledge, in which case the examinee succeeds in classifying

every single option as true or false following independent attempts each of which has a λ probability of success. Total ignorance is represented at the bottom and appears as a result of failure to classify any option, each occurrence of which has the probability, $1 - \lambda$. Between these extremes, there are three rows representing degrees of partial knowledge ranging from knowledge to classify three options to knowledge to classify only one option. Multipliers to the left of the rows indicate the number of different ways that a given configuration of classified and unclassified options can occur at each state (knowledge level).

Only knowledge is involved in determining the final response to an item for the first and second knowledge levels, where the examinee gets the correct response due to the total knowledge or when three out of four options have been classified. However, when fewer than N-1 options have been classified (i.e., the third and fourth knowledge states in the diagram), an assumption about identifiability of distractors, the probability that an answer is among the classified options, has to be made. Such probabilities are stated as x_is at the upper branch of the fifth link in the figure in terms of the probability that the correct answer has been classified. Consequently, the probability that the correct answer has not been classified is $1 - x_i$, as shown at the lower branch of the fifth link. The probability that the correct answer is among the two recognized options is x_1 and the probability that the single recognized option is an answer is x_2 in the diagram. Various assumptions can be made at this point by selecting different values for the x_is. Detailed discussion of these assumptions and the

rationale for each assumption is presented in section 3.1. However, the assumption about identifiability of distractors is not relevant for knowledge level five, where an examinee is totally ignorant and fails to classify any of the options.

When the correct answer is not included in the classified options, an assumption regarding the (individual) examinee's willingness to guess comes into play at the sixth link of the tree diagram. There are three circumstances under which an examinee may guess:

1. when able to classify none of the options (totally ignorant). In this case guessing is assumed to be at random among four options, as shown in the fifth row of the tree diagram.

2. when able to classify one distractor. Then the examinee may guess at random among three options, as shown in the fourth row of the tree diagram.

3. when able to classify two distractors. In this case the examinee may guess at random among two options, as shown in the third row of the tree diagram.

Unlike what was originally proposed in the theory (García-Pérez 1987), namely, a single value γ representing the probability of guessing in all three cases, three different values, γ_0, γ_1, γ_2, are used in this study for the probability that the examinee will actually guess among four, three, or two options as described above. Thus, γ_i is not the probability of guessing correctly, but it is the probability of guessing at all. The examinee's willingness to guess is stated as γ_i at the upper branches of the sixth link of the tree diagram. The probabilities of leaving the item as

an omission 1 - γ_i are stated at the lower branches of the sixth link. Three different assumptions are also made for γ_i. Detailed discussion of these different assumptions and their rationale is presented in section 3.2.

All these variables λ, γ_i, x_i are combined together to develop the finite sate models. The probabilities of all paths leading to the same outcome in the tree diagram (Figure 1) are summed up to obtain the probabilities of response outcomes correct (c) in equation 1, wrong (w) in equation 2, and unanswered (u) in equation 3 as shown in Appendix A. These are theoretical probabilities of each response outcome as given by the finite state model for the item and situation described. These equations also analytically define the theoretical relationships between the value of the latent variable and the probability of each response outcome.

3. The Finite State Models of the Study

In finite state theory, different finite state models can be developed by making different assumptions about identifiability of distractors and guessing strategy which in turn determine the values of x_is and the relationships among γ_is in the tree diagram and equations mentioned in the previous section. The assumptions considered in this study are presented below.

3.1. Assumptions about the identifiability of distractors

The probability that an answer is among the recognized options is represented by the x_is at the fifth link of the tree diagram. The values of x_is depend on the relative difficulty of identifying the correct answer as opposed to the distractors.

García-Pérez (1985) suggested that if all items are well written, examinees should have equal probability of being able to identify the correct option and distractors. That is, the probability of the correct answer being included among 'k' identified options out of 't' given options in the item would be k/t.

Based on this rationale, the previous studies on finite state theory, García-Pérez (1987, 1990), García-Pérez and Frary (1989, 1991a) assumed that for a given examinee, the probability of being able to classify a randomly selected correct option is the same as for a randomly chosen distractor. For this case, it is assumed that $x_1 = 1/2$ and $x_2 = 1/4$. Consequently then $(1 - x_1) = 1/2$, and $(1 - x_2) = 3/4$. Since this assumption is quite demanding in a real situation, two other assumptions are also considered for x_is values in this study. The other possible extreme case would be that whenever examinee classify fewer than N-1 options, these classified options are distractors. For this case, we can assume that $x_1 = x_2 = 0$, and therefore, $(1 - x_1) = (1 - x_2) = 1$. Middle points between this extreme and the assumption of equal identifiability are, $x_1 = 1/4$ and $x_2 = 1/8$. The summary of the three sets of values of x_is to be used at each of the two branching points on the fifth link of the tree diagram are presented in Table 1.

Insert Table 1 about here

3.2 Assumptions about variations in guessing strategy

At the three branching points on the sixth link of the tree diagram, various assumptions concerning γ_i can be made based on the possible guessing behavior under the circumstances. For number-right scoring, it could be assumed that the values of γ_i would be $\gamma_1 = \gamma_2 = \gamma_0 = 1$ because examinees should always guess regardless of how many options they can classify correctly. For formula scoring, a plausible assumption would be $\gamma_1 = \gamma_2 = 1$ and $\gamma_0 = 0$, because examinees are instructed to guess when they are able to classify one or more options or have partial knowledge and to omit the item when they are totally ignorant.

However, when a test is administered with instructions that are ambiguous regarding guessing, an examinee's willingness to guess might vary according to the examinee's partial knowledge or the number of options the examinee is able to classify. For instance, it is not unreasonable to assume that an examinee might have a greater propensity for guessing after classifying two distractors than when only one distractor is classified. Similarly, an examinee might have a greater propensity for guessing after one distractor is classified than when totally ignorant. Based on this rationale, the following three assumptions are made regarding propensity for guessing.

Let γ_0 be the probability that an examinee will guess when no options has been classified.

Let $\gamma_1 = \gamma_0 + y_1(1 - \gamma_0)$ where $0 \leq y_1 \leq 1$ be the probability that an examinee will guess when 1 option has been identified.

Let $\gamma_2 = \gamma_0 + y_2(1 - \gamma_0)$, where $y_1 \leq y_2 \leq 1$ be the probability that an examinee will guess when 2 options have been identified.

These assumptions make $\gamma_2 \geq \gamma_1 \geq \gamma_0$. Four different pairs of values for y_js are selected. These values represent a range of plausible assumptions about guessing strategies under assumptions one through five in the previous section. One possible guessing behavior is that an examinee always guesses when able to classify one or two options and respond according to the value of γ_0 when totally ignorant. For this behavior, $\gamma_1 = \gamma_2 = 1$ or equivalently $y_1 = y_2 = 1$. In this case the value of γ_0 remains an examinee characteristic.

Another possible guessing behavior is that the probability of guessing is the same regardless of the number of options an examinee is able to classify. This behavior was described as random omission behavior in García-Pérez and Frary (1989). For this behavior $\gamma_0 = \gamma_1 = \gamma_2$ or equivalently $y_1 = y_2 = 0$, and all these three γ_is are determined by the examinee's value for γ_0. In between these two extremes two other pairs of values for (y_1, y_2), namely, (1/2, 2/3), (1/4, 1/3) are selected to vary the relations among the γ_is. Table 2 summarizes the selected values of y_js for the different assumptions concerning guessing strategy.

Insert Table 2 about here

3.3 Models to be considered

The three different possibilities about identifiability of distractors as shown in Table 1 may be combined with the four different guessing strategies as shown in Table 2 to yield 12 different models as shown in Table 3. Assumptions 1-3 about the x_is and assumptions 1-4 about the y_js will be referred to by the subscripts k (k=1, 2, 3) and m (m=1, 2, 3, 4) as follow:

M_{km} refers to the model incorporating assumption *k* about the x_is and assumption *m* about the y_js.

$M_{k.}$ refers to all models incorporating assumption *k* about the x_is

$M_{.m}$ refers to all models incorporating assumptions *m* about the y_js.

Insert Table 3 about here

---

4. Estimation of the parameters

The equations for c, w, u, from section 2 are combined with the choice of the values of x_i, y_j, representing assumptions, to define relationships between observable response outcomes (the proportions of correct, incorrect, and unanswered items) and the unobservable λs and γs. The equation resulting from solving equation 1 and equation 2 simultaneously provides the solution for λ and is referred to as finite state scoring polynomial.

Different numerical methods have been used to estimate examinees' ability parameter (λs) as the circumstances allowed. Both García-Pérez (1987) and García-Pérez and Frary (1989) estimated λ by finding the root or solution of finite state scoring polynomials of degree four and six for the three-option multiple-choice test and four-option multiple-choice test with different response behaviors under the conventional administration mode.

In García-Pérez (1990), each examinee's ability λ was estimated by using the minimum chi-square method for five-option multiple-choice items under answer-until-correct administration.

García-Pérez and Frary (1991a) studied five-option multiple-choice items each consisting of four content options and the option, "none of the above", with the instruction to attempt every item. Least squares methods were used to estimate ability in this study.

García-Pérez (1991), using simulation, compared the performance of four-option multiple-choice items consisting of three content options and a "none of the above" option, with conventional four-choice items under formula scoring instructions. He used maximum likelihood procedures to find the point and interval estimates of each examinee's λ.

An estimate of λ, based on a nonlinear function of the proportions of multiple-choice responses that fall into the possible response categories under the conditions stated for each scoring model, is referred to as a finite state score. For a conventional

multiple-choice test, the possible categories are correct, incorrect, and unanswered. If the test consists of a random sample of all available items, then a finite state score will estimate the proportion of options known in the population of available options. Hence, λ is directly interpretable as a measure of the level of knowledge the examinee has.

There are two main advantages to finite state scores:

1. The model used to produce them can be tailored to a number of circumstances surrounding test administration including the response behavior of the examinees. Finite state theory models the response behavior appropriately by establishing the contribution of each knowledge state to the probability of each possible response outcome. Provided that the number of independent response categories exceeds the number of parameters to be estimated, this behavior can be inferred by testing the fit of a spectrum of models constructed by systematically varying underlying assumptions and selecting the model that fits best.

2. The metric of finite state score is directly interpretable as representing the proportion of options that the examinee knows. Thus, comparisons among examinees or groups of examinees are made on an absolute rather than a relative basis (see García-Pérez & Frary, 1989). Even if different finite state models are required for two groups because their response behavior differs, their scores can be compared on the same metric.

Although there are two important parameters in the finite state theory, previous studies were mainly concerned with the ability parameter (λ). In García-Pérez (1990, 1991) and García-Pérez and Frary (1991a), the parameter representing an examinee's willingness to guess (γ) was not relevant because these studies were conducted for tests which required an answer to every item. Although γ could be obtained in the studies of García-Pérez (1987) and García-Pérez and Frary (1989), it was not calculated or discussed in detail. In this study, γ was modified as explained in section 3.2, and calculated to explore the relationship between γ and other selected variables.

The responses from four-option multiple-choice items administered under the conventional response mode were used in this study. These responses provided three response categories, correct, incorrect, and unanswered, to estimate examinee parameters λ and γ. Thus, the number of independent response categories did not exceed the number of parameters to be estimated and therefore, it is not possible to conduct a goodness of fit study. Although, the data did not have enough information to carry out the goodness of fit test between the models and data, they provided an occasion for studying the consequences of the adoption of various assumptions on estimates, to gain better understanding of variables in the finite state models. Hence, examinee parameters λ and γ were estimated in each model, as if all examinees in the sample had uniformly followed the test-taking behavior assumed in the model.

An examinee's ability parameter λ was estimated by finding the root of a sixth degree finite state scoring polynomial. The observed values of number of right

answers, number of wrong answers, and number of unanswered items for each examinee were expressed as proportions, and referred to as *C*, *W*, *U* respectively. These values *C*, *W*, *U* were taken as estimates of the theoretical probabilities *c*, *w*, *u*, in equations 1-3 of Appendix A. Equations 1 and 2 were expanded and combined to yield a finite state scoring polynomial, equation 4, from which the examinees' finite state scores (λs) were calculated. Equation 5 was derived from equation 2 to estimate the examinees' willingness to guess (γ_0s). Equations 4 and 5 were used in a FORTRAN program, (see Appendix B) to calculate twelve pairs of λs and γ_0s for each examinee according to the observed values of *C*, and *W*. This program was adapted from a BASIC program of García-Pérez's (1987). A detailed explanation of the operation of the program is given in Appendix C. Although only γ_0 was calculated from equation 5 as required for later analyses, the values of γ_1 and γ_2 could be calculated from γ_0 and the corresponding values of y_j.

CHAPTER III

Data and Methods of Investigation

1. Data Source

Data for this study came from the spring, 1989, administration of the Test of Achievement and Proficiency (TAP, Riverside Publishing Company, 1986) to about 65,000 Virginia public school eleventh grade students. Administered statewide under the sponsorship of the Virginia Department of Education, this battery contained six subtests, namely, reading comprehension, mathematics, written expressions, using sources of information, social studies and science. In addition, 15 survey items explored the examinees' academic backgrounds and attitudes. Gender and ethnicity of the examinees were also recorded. Number-right scores on each subtest were reported to the schools and students to aid in a variety of educational decision making processes.

At the beginning of the test booklet of TAP (Scannell, 1986, p. 2) the test instructions are presented as follows (underlining added):

> EARNING YOUR BEST SCORE
>
> Some students receive lower scores on tests than they could receive, simply because they do not take the tests in the most efficient manner. The information below is provided to help you earn your best score. As you take the test, remember these points

1. If you are not absolutely sure about the answer to a question, but think you know the correct answer, mark a choice. You will earn your best score if you attempt all questions for which you think you know the answers. You will not lose any points for incorrect choices.

2. There are some questions on each test which you may not be able to answer. Do not linger over difficult questions; omit these and go on to easier ones. You may return to omitted questions at the end of the test if there is time remaining.

This underlined part of the instructions fails to suggest the best test taking strategy for number-right scoring. (All TAP tests were scored number-right thus making any omissions counterproductive from an examinee standpoint.) Because of the deficient instructions or other reasons, there were many omitted responses. All omissions were assigned to two categories, namely, embedded omissions and trailing omissions. Embedded omissions were defined as omissions prior to the last answered item. Trailing omissions (or unreached items) were defined as omissions occurring after the last answered item.

Responses to the science subtest of the TAP were used for this study. This test consists of 54 four-choice items, all of whose options were content options (i.e., none of the above was not a response choice). The science subtest was chosen because its responses include a relatively large number of embedded omissions and relatively few trailing omissions.

Among the examinees with omissions, 5,287 had embedded omissions and no trailing omissions. These examinees were selected for this study because the occurrence of embedded omissions guarantees that these students did not employ the optimal number-right test response behavior yet had time to reach the end of the test. Examinees with trailing omissions may have been influenced by speededness, a factor beyond the scope of this study. Hence, examinees with trailing omissions were not used in the analysis. The range of embedded omissions of the sample was from one (2%) to twenty (37%). Nineteen examinees who had more than twenty embedded omissions were also excluded because factors other than knowledge and test-taking behaviors, such as lack of motivation may have caused such excessive numbers of omissions.

2. Comparisons of scores

The comparison of finite state scores and number-right scores was aimed at determining to what extent finite state scoring might lead to conclusions that were different from those reached when number-right scoring was used. Throughout the analysis number-right scores were expressed as proportions of correct answers and rescaled with the linear transformation $RC = 4C/162 - 1/3$, where RC is a rescaled number-right scores and C is a number-right score. The transformation rescaled linearly with 13.5 or less (guessing level) corresponding to 0 and 54 (perfect score) corresponding to 1. The transformation put the number-right scores on the same scale

as the finite state scores. *RC* values of 154 examinees were less than 0 and rounded to 0.

The comparisons among the finite state scores calculated from different models were also considered to examine the extent to which finite state models differ in their ability estimations. Both comparisons were made based on the three different aspects of scores: (1) ranking characteristics, (2) estimation of examinee's ability, and (c) effect on conclusions based on group differences.

2.1 Comparisons of ranking characteristic

Ranking characteristic of the scores pertain to the use of test scores to rank order the examinees according to their performance on the test relative to other examinees. The examinees' ranks resulting from number-right scores were compared with those from finite state scores from each model. Spearman rank correlations between number-right scores and finite state scores from each model were calculated for this purpose. Moreover, to examine whether finite state scores from different models provide similar rank orders of examinees or not, Spearman rank correlations between finite state scores from all pairs of models were also calculated.

2.2 Comparisons of ability estimates

Although a major use of test scores in current testing practice is the ranking of examinees according to their relative performances on the test, the potential of test scores to estimate an examinee's level of knowledge is also important. Number-right scores usually provide only ranking information whereas finite state scores provide

estimation of the examinees' levels of knowledge in addition to providing rankings. However, it is essential to find out how the estimates of the examinees' ability differ across the 12 finite state models, because each model incorporates different plausible assumptions concerning identifiability of distractors and guessing strategy.

Therefore, descriptive statistics for the finite state scores from each model were calculated to examine how these plausible assumptions affected the estimated ability. Descriptive statistics for the number-right scores and rescaled number-right scores were also calculated and compared with those of finite state scores. This was done because number-right scores, expressed as percentages, are often interpreted as percentages of some body of knowledge known (i.e., ability estimates).

2.3 Comparisons of group differences revealed by different type of scores

Test scores for schools and school systems are increasingly subject to public disclosure. Test scores have been used and misused as indicators of achievement levels of individual examinees, of schools and school systems, of the effectiveness of programs, and of the quality of teachers. Group comparisons are often made, discussed, and evaluated based on test scores. For this reason, conclusions about group differences based on scores were investigated in this study to examine whether groups established according to apparent characteristics seem to differ more or less depending on which set of scores were considered. Of particular interest were the differences between group mean scores and the rank order of group means. The analysis was conducted to examine whether these two aspects of group comparisons

were consistent across 13 different scores, namely, 12 finite state scores and rescaled number-right scores.

Due to the concern about possible bias against ethnic minorities and females, race and gender were selected as grouping variables. Moreover, to explore the consequences of various finite state models on examinees' ability levels, self-reported grade-point-average (GPA) from the past three years, and years of high school science course work were also selected as grouping variables. Originally, GPA was collected on a five-point scale and years of high school science course work was collected on an eight-point scale in the survey questionnaire. However, some categories of these two variables were collapsed to make four categories for each variable in the analyses. There were also large number of missing values (1778 for GPA and 1664 for years of high school science course work) on these variables.

'Omissiveness' was defined as the ratio of the number of embedded omissions to the number of wrong answers. Low values of omissiveness suggest few omissions and many wrong answers (high willingness to guess), whereas high values of omissiveness suggest many omissions and few wrong answers (low willingness to guess). The value of omissiveness for each examinee was calculated according to the definition and rounded to two decimal places. Although the range of omissiveness was from 0.02 to 4.00, examinees were not evenly spread across the range. At the bottom end, there was a heavy concentration (n= 2588) between 0.02 and 0.07, but there were only five examinees above 2.00 at the other end.

Given the range and concentration of omissiveness, it seemed reasonable to split the sample into three distinct groups. The least omissive group consists of examinees whose omissiveness value is less than 0.07, the moderate omissive group consists of examinees whose omissiveness value is between 0.07 and 0.47, and the most omissive group consists of examinees whose omissiveness values is greater than 0.52. Three hundred three examinees who had omissiveness values of 0.07, and 60 examinees who had omisssiveness values between 0.47 and 0.52 were left out to make the groups more distinct from each other. To determine the effect of different finite state models on examinees with various levels of omissiveness, score comparisons were also made for the omissiveness levels.

The statistical design selected to test the group differences based on type of scoring was a two-factor fixed design, analysis of variance with repeated measures on one factor (Ott, 1989, chap. 17). The two factors were group membership and type of scoring. Repeated measures were scores from 13 types of scoring. Series of repeated measures analysis of variance were made for 13 types of scoring with five selected grouping variables, gender, race, GPA, science course work, and omissiveness. Significance levels associated with Huynh-Feldt adjusted F tests (SAS Institute, Inc. 1989) were used for the conclusions of all statistical tests to adjust for dependence among the repeated measurements. The significant interaction effects between type of scoring and group membership were also examined graphically.

3. Exploration on examinees' estimated γ_0

Three aspects of examinees' estimated willingness to guess were investigated: (1) comparisons of estimated γ_0 from 12 finite state models, (2) comparisons of group differences based on estimated γ_0 revealed by different models, and (3) relationships between estimated γ_0 and other selected variables.

3.1 Comparisons of estimated γ_0 from different models

Descriptive statistics of γ_0 were calculated to compare the estimated willingness to guess across the 12 models. The descriptive statistics were examined to assess how the variations in the assumptions about identifiability of distractors (x_is) and guessing strategies y_js affected the estimated willingness to guess.

3.2 Comparisons of group differences based on estimated γ_0 revealed by different models

A two-factor fixed design, analysis of variance with repeated measures on one factor (Ott, 1989, chap. 17) was selected to test the group differences based on estimated γ_0 across the 12 finite state models. The two factors were group membership and models. Repeated measures were estimated γ_0 from 12 models. Grouping variables were the same as those from the comparisons of scores, namely, gender, ethnicity, GPA, years of high school science course work, and omissiveness. Series of repeated measures analysis of variance were made for five grouping variables with 12 finite state models. Significant levels associated with Huynh-Feldt adjusted *F* tests (SAS Institute, Inc. 1989) were used for the conclusions of all statistical tests.

The significant interaction effects between models and group membership were also examined graphically.

3.3 Relationships between estimated γ_0 and other selected variables

Theoretically, γ_0 is a quantitative estimate of an examinee's willingness to guess when the examinee is totally ignorant. One of the possible criteria to evaluate estimated γ_0 is the relationship between estimated γ_0s and other possible indicators of examinee's willingness to guess such as number of embedded omissions and omissiveness. Therefore, Pearson correlations between number of embedded omissions, omissiveness, and estimated γ_0 were calculated. These correlations were expected to be strong in the negative direction, according to the definitions of variables. Furthermore, Pearson correlations between finite state scores, number-right scores and estimated γ_0 were calculated to investigate how examinees' estimated willingness to guess related to examinees' estimated ability.

CHAPTER IV

Results

Results from all analysis are organized into two main sections, findings on scores and findings on estimated willingness to guess. In each section, the results are reported according to the sequence of research questions listed in Chapter 1, (P 7-8).

1. Findings on scores

1.1 Comparisons of ranking characteristics

Table 4 gives Spearman rank correlations between number-right scores and finite state scores from each model.

Insert Table 4 about here

Spearman rank correlations among finite state scores from different models are not reported as a table, because they all are perfect or nearly perfect (.998 and above).

1.2 Comparisons of ability estimates

In order to compare the ability estimates from different types of scoring, means, medians, and standard deviations for 14 different scores, 12 finite state scores, number-right scores, and rescaled number-right scores, are reported in Table 5.

Insert Table 5 about here

The mean of number-right scores, expressed as proportions of correct answers (0.54) is higher than the mean of finite state scores from all 12 models, whereas the mean of rescaled number-right scores (0.38) is closer to the mean of finite state scores from models $M_{2.}$ than those of other models. The mean finite state scores differ greatly across models, ranging from a mean of 0.31 to 0.44. Specifically, there is a considerable change in the magnitude of means from the models with different assumptions about the identifiability of distractors x_is. The order of models according to the magnitude of mean finite state scores is $M_{1.} < M_{2.} < M_{3.}$. However, there is not much difference among the standard deviations of finite state scores across 12 models. All standard deviations are about 0.16 or 0.17.

According to the definition of finite state scores, Table 5 shows that, on the average, an examinee would be able to classify (correctly) approximately, 31% of options if models $M_{1.}$ are applied, 38% of options if models $M_{2.}$ are applied, and 44% of options if models $M_{3.}$ are applied.

The findings in Table 5 permit another important observation of finite state models: if attention is shifted from assumptions about x_is to assumptions about the y_js, then very little difference (difference at the third decimal place) is found among the means of models incorporating the same assumption about x_i with different assumptions about guessing strategy ($M_{k.}$). Mean finite state scores from the 12 finite

state models are plotted against four guessing strategies in Figure 2 to show the effect of variations of assumptions on ability estimates.

Insert Figure 2 about here

In the figure, the lines connect four mean finite state scores from $M_{k.}$ (models incorporating the same assumption about identifiability of distractors). These lines are nearly parallel across the four guessing strategies.

1.3 Comparisons of group differences revealed by different scores

Sample size of the groups, means and standard deviations of 13 different types of scores; 12 finite state scores and rescaled number-right scores, are reported in tables for each grouping variable as follows:

Table 6, means and standard deviations by gender;

Table 7, means and standard deviations of scores by ethnic group;

Table 8, means and standard deviations of scores by GPA level;

Table 9, means and standard deviations of scores by years of science course work; and

Table 10, means and standard deviations of scores by omissiveness level.

Insert Tables 6, 7, 8, 9, and 10 about here

The major concern of group comparisons across different types of scoring is to examine how the different types of scoring impact on examinees with different characteristics. Specifically, is there consistency in the order of group means across different types of scoring and consistency in the differences between group means across different types of scoring? Therefore, the results of repeated measures analysis of variance for all group comparisons are presented in Tables 11, 12, 13, 14, and 15.

Insert Tables 11, 12, 13, 14, and 15 about here

As shown in tables (Tables 6 to 10), the sizes of subgroups ranges from 2744 to 74. Due to the large sample sizes, the results of statistical tests on all interaction effects are highly significant ($p < .0001$), although the differences among means are found at the third decimal place for some variables.

Hence, mean scores versus group membership are plotted for each group comparison to look for meaningful interactions, i.e., the interaction effects that are both statistically significant and are large enough to be meaningful. The plots from all five group comparisons are presented in Figures 3, 4, 5, 6, and 7.

Insert Figures 3, 4, 5, 6, and 7 about here

One percent or more change in the differences between mean scores from two groups were interpreted as meaningful interaction. This criterion was selected

because, on a 54 item test with four-choices per item, 1% of all 216 options is more than two options correctly identified using finite state scoring. The meaningful interactions are found only in the comparisons of gender, GPA, and omissiveness.

2. Findings on estimated willingness to guess: γ_0

2.1 Comparison of estimated willingness to guess (γ_0) from 12 finite state models

Table 16 provides the medians and interquartile ranges of the distributions of estimated γ_0 from 12 finite state models. Since all distributions of estimated γ_0 are negatively skewed, medians and interquartile ranges instead of means and standard deviations are compared across the models.

Insert Table 16 about here

The models incorporating assumptions about guessing strategy 1, $M_{.1}$, have the lowest medians (0.85 for M_{11}, 0.78 for M_{21}, 0.65 for M_{31}) with large interquartile ranges (0.31, 0.51, and 0.83 respectively) compared to those from other finite state models. Models incorporating other assumptions about guessing strategies 2, 3, and 4 ($M_{.2}$, $M_{.3}$, and $M_{.4}$) have median above 0.90 and smaller interquartile ranges regardless of which assumption about identifiability of distractors x_is is used. The medians of estimated γ_0 are approximately 0.92 for models $M_{.2}$, 0.95 for models $M_{.3}$, 0.96 for $M_{.4}$ and interquartile ranges are about 0.12 for models $M_{.2}$, 0.07 for $M_{.3}$ and 0.06 for models $M_{.4}$.

The plot of medians of estimated γ_0 in Figure 8 also shows that assumptions of x_is do not contribute much to the estimation of γ_0, when models $M_{.2}$, $M_{.3}$ and $M_{.4}$ are applied. However, when models $M_{.1}$ are applied, estimated willingness to guess is affected by the assumptions about x_is.

Insert Figure 8 about here

2.2 Comparisons of group differences based on estimated γ_0 across different models

Group differences based on estimated examinees' willingness to guess are also compared across the models to examine the impact of variations of assumptions on group differences. Sample sizes, means and standard deviations of examinees' estimated γ_0 for each group on all five selected variables are reported in the tables as follows: gender in Table 17, ethnic group in Table 18, GPA level in Table 19, years of science course work in Table 20, and omissiveness level in Table 21.

Insert Tables 17, 18, 19, 20, and 21 about here

The results of repeated measures analysis of variance on estimated γ_0 for all five grouping variables are reported in Tables 22, 23, 24, 25, and 26 as the same order of the Tables 17 to 21.

Insert Tables 22, 23, 24, 25 and 26 about here

Due to the large sample size, interaction effects between different models and group membership on estimated γ_0 are statistically significant for all group comparisons except the comparison for gender. The plots of mean estimated willingness to guess versus group membership for the group comparisons with significant interaction are presented as follows: Figure 9 for ethnic groups, Figure 10 for GPA, Figure 11 for years of science course work, and Figure 12 for omissiveness level.

Insert Figures 9, 10, 11, and 12 about here

Based on the visual display of these figures, more than 4% change in the differences between mean estimated willingness to guess from two groups were interpreted as meaningful. Four meaningful interactions were found for the grouping variables ethnicity, GPA, years of science course work, and omissiveness.

2.3 Findings on relationships between estimated γ_0 and other selected variables

The correlations between number of embedded omissions, omissiveness, and examinee estimated γ_0 are reported in Table 27.

Insert Table 27 about here

As expected, estimated willingness to guess has moderate to high negative correlations with both number of embedded omissions and omissiveness. Although the distributions of estimated γ_0 are negatively skewed, the examination of scatter plots suggests that there is no nonlinearity in the relationship. The magnitude of correlations are noticeably higher for the correlations between estimated γ_0 and omissiveness than those between estimated γ_0 and number of embedded omissions.

The willingness to guess estimated from models $M_{.1}$ have the lowest correlation with both number of embedded omissions and omissiveness. However, γ_0 estimated from the models incorporating assumptions 2, 3, and 4 about guessing strategy ($M_{.2}$, $M_{.3}$, $M_{.4}$) have very high negative correlations (above 0.9) with omissiveness.

Insert Table 28 about here

Table 28 gives the correlations between estimated ability, finite state scores, number-right scores and estimated γ_0. Low negative correlations are found between ability estimates and estimated γ_0 for all models except for models $M_{.1}$. Models $M_{.1}$ have moderately high negative correlations, approximately -.7, between ability estimates and willingness to guess estimates.

CHAPTER V

Discussion and Conclusions

The major purpose of this study was to explore the effects of different assumptions about identifiability of distractors and about guessing strategy on ability estimates and estimated willingness to guess. Hence, these effects are discussed under two separate sections as follows.

1. Ability estimates

1.1 Comparisons of ranking characteristic

Very high (nearly perfect) Spearman rank correlations between number-right scores and finite state scores from each model (Table 4) led to the conclusion that finite state scores provide essentially the same rankings of examinees as number-right scores regardless of which finite state model is considered. Moreover, perfect or nearly perfect Spearman rank correlations among finite state scores from different models also show that the various assumptions concerning identifiability of distractors and guessing strategies do not affect the rank order of examinees.

In any case, so long as it is assumed that all examinees behave the same, highly similar rankings occur from any two finite state models, even though, obviously, this (supposedly uniform) behavior does not conform uniformly well to all 12 models. Moreover, whatever this behavior is, it certainly does not conform to that assumed for number-right scoring, due to the presence of omissions. Yet the ranking provided by number-right scoring is virtually indistinguishable from that produced by

any finite state model. Therefore, it is concluded that insofar as ranking is concerned, conformity of behavior to model assumptions has little or no effect, and finite state scoring has no benefit over number-right scores.

1.2 Comparisons of ability estimates

The figures in Table 5, descriptive statistics of 14 different scores, are quite informative. These figures show that on the average, rescaled number-right scores are about the same as ability estimates from models $M_{2.}$. Findings on comparisons among finite state scores from the 12 adopted models show that on the average, ability is the highest when models $M_{3.}$, models incorporating assumption 3 concerning identifiability of distractors, are applied. On the other hand, estimated ability is the lowest when models $M_{1.}$, models incorporating assumption 1 concerning identifiability of distractors, are applied.

As defined in chapter 2 (p. 13), the assumption concerning identifiability of distractors is actually an assumption as to the probability that the correct answer is among the classified options if an examinee is able to classify fewer than N-1 options. In terms of that definition, estimated ability is the highest when the probability that the correct answer is among the classified options (if an examinee is able to classify fewer than N-1 options), x_i, is assumed to be 0. On the other hand, estimated ability is the lowest when the probability that the correct answer is among the classified options (if an examinee is able to classify fewer than N-1 options), x_i, is assumed to be the ratio of classified options to the total number of options per item.

A possible explanation for this outcome arises from the tree diagram (Figure 1). There it may be noted that, if x_is are assumed to be 0, two of the paths leading to the response outcome *c* will disappear (at knowledge level 3 and 4). Then a higher λ is required to account for the same proportion of correct responses due to this reduction of potential paths. In other words, only a higher λ will account for the same proportion of correct answers via fewer paths. Further investigation of the identifiability of distractor, variable (x_i) is needed.

The magnitudes of mean finite state scores change considerably across the models M_k. (models incorporating different assumptions about identifiability of distractors x_i), but they change very slightly (at the third decimal place) across the models incorporating the same assumption about x_i but with different assumptions about guessing strategy. Thus, assumptions about guessing strategy have surprisingly little effect on ability estimates. Figure 2, the plot of mean finite state scores from the 12 models also reveals this conclusion graphically.

1.3 Comparisons of group differences revealed by different type of scores

In view of the results shown in Table 5 and Figure 2, large main effects of the repeated measures factor in the analysis of variances were expected. These will not be discussed further here. Instead the discussion is limited to the interactions that analyses were meant to investigate. The significant interaction between gender and scoring model is not strong. Table 6 and Figure 3 both report the comparison of scores by gender and show that when models M_3. (models incorporating the

assumption 3 concerning identifiability of distractors) are applied, the differences between gender groups are smaller than those when other models and number-right scoring are applied. This finding shows that depending on which model fits the behaviors of (all of) the examinees, there may or may not be actual gender differences in ability.

The comparison of scores for the GPA (Table 8, Figure 5) reveals an interesting outcome with respect to the difference between finite state scores and number-right scores. For low GPA groups (i.e., GPA C+ and C- groups), average number-right scores are the same or slightly (1%) lower than mean finite state scores from models $M_{2.}$. However, for high GPA groups (i.e., GPA A and B groups), average rescaled number-right scores are somewhat higher than average finite state scores from the models $M_{2.}$. The difference is more obvious for the highest GPA group. A similar pattern can be seen in the comparison of mean scores from groups defined by years of science course work (Table 9, Figure 6) although the difference is not as pronounced as in GPA groups. Both these two interactions reflect the nonlinear relationship between number-right scores and finite state scores.

Compared to the above two interactions, a stronger interaction is found in the comparison of scores for the groups defined by omissiveness level. Table 10 and Figure 7 show that examinees from moderate and most omissive levels are found to have higher ability estimates than the least omissive level. Examinees from the least

omissiveness level seem to make more incorrect guesses compared to those who guess less often. Thus, the ability estimates are lower for the least omissive examinees.

The more interesting finding concerns the differences between the moderate omissive level and the most omissive level. Not only is the difference between these two groups relatively larger in number-right scoring than in finite state scoring (about 6% in number-right scoring and about 1% or less in all finite state scoring), but also the rank order of the group means are different. This finding suggests that finite state scores may have the potential to compensate for the differences in the examinees' omissiveness in the production of ability estimates.

2. Estimated willingness to guess

2.1 Comparisons of estimated γ_0 from 12 finite state models

Table 16 and Figure 8, descriptive statistics for estimated willingness to guess (γ_0), show that when models $M_{.1}$ (models incorporating the assumption 1 concerning guessing strategy, i.e., models assume that examinees always guess when they have partial knowledge and respond according to the value of γ_0 when they are totally ignorant) are applied, the average estimated γ_0 is lower than when models incorporating other assumptions about guessing strategy are applied. This finding is logical, because guessing strategy 1 assumes that examinees always guess when they can eliminate one or more distractors. Thus, many of the wrong answers must result from guessing with partial knowledge and few from guessing with total ignorance. So, when examinees are ignorant, they would guess less often. This finding shows

that the extent of estimated examinee willingness to guess at the state of total ignorance differs according to the guessing strategy assumed.

Moreover, when models $M_{.2}$, $M_{.3}$, $M_{.4}$ (models incorporating assumptions 2, 3, and 4 of guessing strategy) are applied, the average estimated γ_0 is above 0.9 regardless of which assumption about identifiability of distractors is considered. Therefore, it is concluded that the assumptions about identifiability of distractors x_i have little effect on γ_0 except in the case of models $M_{.1}$.

Table 16 and Figure 8 also show that the average γ_0 also changes considerably among the models $M_{.1}$ (models incorporating the same assumption 1 concerning guessing strategy with different assumptions about identifiability of distractors). When model M_{11} (the model incorporating the assumption 1 concerning guessing strategy with the assumption 1 concerning identifiability of distractors) is applied, the average γ_0 is the highest among the models $M_{.1}$ and closer to those when other models with different assumptions about guessing strategy ($M_{.2}$, $M_{.3}$, $M_{.4}$) are applied.

On the other hand, when model M_{31} (the model incorporating assumption 1 concerning guessing strategy with assumption 3 concerning identifiability of distactors) is applied, the average estimated γ_0 is the lowest among the models $M_{.1}$ and much lower than when other models with different assumptions about guessing strategy are applied. Therefore, it is concluded that when models $M_{.1}$ are applied, the estimated γ_0 is affected by the assumptions about identifiability of distractors.

This finding can be explained from the tree diagram (Figure 1). Consider the fixed number of wrong answers for any given examinee. Some of these resulted from items for which the examinee's knowledge is level 3, some from level 4, and some from level 5. The assumption about guessing strategy 1 assumes $\gamma_1 = \gamma_2 = 1$. For this given situation, the following consequences occurr at knowledge level 3, 4, and 5. At knowledge level 3:

under assumption 1 about identifiability of distractors ($x_1 = 1/2$, $x_2 = 1/4$), the probability of getting a correct answer, $P(c)_1$ would be $1/2 + 1/4 = 3/4$.

under assumption 3 about identifiability of distractors ($x_1 = x_2 = 0$), $P(c)_3$ would be 1/2. Thus, $P(c)_1 > P(c)_3$.

As a result, the probability of getting a wrong answer resulted from assumption 1 about identifiability of distractors, $P(w)_1$ is less than the probability of getting a wrong answer resulting from assumption 3 about identifiability of distractors $P(w)_3$. That is, $P(w)_1 < P(w)_3$.

At knowledge level 4:

$$P(c)_1 = 1/4 + 3/4\ (1/3) = 1/2$$

$$P(c)_3 = 1/3$$

Thus, $P(c)_1 > P(c)_3$ and $P(w)_1 < P(w)_3$.

Since the number of wrong answers is fixed, it must be true that $P(w)_1 > P(w)_3$ at knowledge level 5. Therefore, an examinee is more likely to have more wrong answers under assumption 1 about identifiability of distractors than under

assumption 3 about identifiability of distractors at knowledge level 5. These wrong answers must have resulted from guessing, because misinformation is not considered in this study. Hence, average estimated γ_0 from M_{11} is bigger than that from M_{31} .

2.2 Comparisons of group differences based on γ_0 revealed by different models

In view of the results shown in Table 16 and Figure 8, the effect of repeated measures was expected to be significant, because of the considerable difference between the estimated γ_0 when models $M_{.1}$ are applied and when other models are applied. However, in the comparison of estimated γ_0 from omissiveness levels, the effect of group membership is bigger than the model effect due to the large differences among the group means of estimated γ_0.

The means of estimated γ_0 from omissiveness levels (Table 21) are also noteworthy. The values of examinee estimated γ_0 from the most omissiveness level are consistently lower than those from the other two levels across the 12 models. This implies that compared to other examinees, examinees from the most omissiveness level are always found to have the lowest estimated γ_0, regardless of which guessing strategy is considered. Since these examinees have high ability estimates, they should know that guessing is beneficial for number-right scoring despite the defective test direction. Nevertheless, they have the lowest estimated willingness to guess. This evidence suggests that guessing behavior is influenced by other factors besides test directions.

The interaction effects between group membership and models with respect to estimated γ_0 reveal that finite state models incorporating different assumptions about guessing strategy have different impact on the estimated γ_0 of the groups of examinees with different characteristics, specifically, ethnicity, GPA, years of science course work, and omissiveness.

All of these interactions involve models with guessing strategy 1 (M_{11}, M_{21}, and M_{31}). In all cases, the pattern of interaction is similar except for the interaction involving omissiveness levels. The difference between the group with the lowest estimated γ_0, and other groups were found to be larger in models $M_{.1}$ than in the other models for groups with relatively lower average estimated abilities, such as the American Indian and black groups as shown in Figure 9, the GPA C-, C+ groups in Figure 10, and the 2yr-, 2yr+ years of science course work groups in Figure 11. This result is consistent with other findings that when models $M_{.1}$ are applied, negative correlations between ability estimates and estimated γ_0 are higher than those when other models are applied. Stronger interaction effects are found in both models $M_{.1}$ and models $M_{.2}$ when estimated γ_0 is compared across omissiveness groups (see Figure 12).

The group comparison findings concerning estimated γ_0 reveal that, depending on which model fits the behavior of (all of) the examinees, the size of the differences between the average estimated γ_0 for any two groups is varied from essentially no difference to a large difference. However, number-right scoring always assumes that

the extent of willingness to guess is the same for every examinee. It should be noted that number-right scoring uses a stronger assumption about examinee willingness to guess than any of the finite state models used in this study. Specifically, number-right scoring assumes that every examinee not only follows the same guessing strategy, but also never fails to guess when the answer is not known.

2.3 Relationships between estimated γ_0 and other selected variables

The findings on relationships between γ_0 and other selected variables provide some information about examinees' willingness to guess. The high negative correlations (above 0.9) between omissiveness and estimated γ_0 reported in Table 27, suggest that omissiveness could be an indicator of examinees' γ_0 when models $M_{.2}$, $M_{.3}$, $M_{.4}$ are applied. Number of embedded omissions could be an indicator of examinees γ_0 when models $M_{.3}$, $M_{.4}$ are applied. Compared to the number of embedded omissions, the slightly stronger correlation of omissiveness indicated that it may be more reflective of examinees' γ_0.

Low to moderate negative correlations between estimated ability, number-right scores, finite state scores and estimated γ_0, except for the models $M_{.1}$ (Table 28), show that high ability examinees are less likely to guess. The findings are consistent throughout the analysis to the effect that the examinees with the highest measures on possibly ability-related variables, namely, GPA and years of science course work, have the lowest γ_0 in all 12 finite state models. The findings on omissiveness levels

also show that the most omissive groups have the highest ability estimates with the lowest willingness to guess estimates.

Moreover, moderately high negative correlations (about -.7) between ability estimates and estimated willingness to guess in models $M_{.1}$ (models with guessing strategy 1 that assumes examinees always guess when examinees have partial knowledge and respond according to the value of γ_0 when they are totally ignorant), imply that high ability examinees are less likely to guess at the state of total ignorance under the assumption that they always guess when they have partial knowledge. This result was anticipated because the guessing behavior of high ability examinees is more likely to come from the states of partial knowledge than the state of total ignorance.

3. Concluding remarks

All of the above discussions and conclusions are made based on findings that resulted from the strong assumption that all examinees in the sample uniformly follow the response behavior assumed in the model. If some examinees follow the behavior assumed in one model and some examinees follow the behavior assumed in another model, the results and conclusions may be different from the above discussion.

These consequences can be illustrated by the following example. In Chapter 1 and 2, it was mentioned that goodness of fit test could not be done for this data. If such a study could have been carried out for this data set, and responses of half of the examinees had been found to fit model M_{11} and responses of the other half had been found to fit model M_{31}. As a result, estimated ability of examinees from the first half

would be finite state scores from model M_{11} and estimated ability of examinees from the other half would be finite state scores from model M_{31}. The rank order correlations between finite state scores and number-right scores in this case is 0.91. Therefore, there is a slight tendency for different ranks to be given to the same examinees, based on finite state scores and based on number-right scores.

This possibility and two major findings of this study (the interaction effects between models and group membership on both ability estimates and estimated willingness to guess, the effect of variation in the assumptions about identifiability of distractors x_is on finite state scores) attest to the need for a goodness of fit study in order to calculate finite state scores that represent fair and accurate estimates of examinees' abilities, i.e., to use the best fitting model.

Due to the impossibility of a goodness of fit test between models and examinees' responses, the proper application of finite state scoring in this study is not possible. The use of finite state scores may be feasible, if the mechanism of the goodness of fit test for common four-choice multiple-choice items with the conventional administration mode is established in future.

However, finite state scoring might reasonably be applicable to multiple-choice items administered with a nonconventional response mode such as answer-until-correct or to multiple-choice items with a noncontent option (e.g., "none of the above"). Each these cases provides more independent response categories than the number of parameters to be estimated. Another possibility for attaining this goal is the use of a

test that consists of items with differing numbers of options, such as four-option multiple-choice items and three-option multiple-choice items (M. A. García-Pérez, personal communication, April 8, 1992).

Willingness to guess estimates from the application of finite state scoring would be useful if one needs to identify examinees who are less or more willing to guess. In fact, finite state theory provides an efficient and flexible approach to explaining how an examinee's ability, partial information, guessing strategy, and item characteristics such as number and type of options, play roles in determining the response outcomes. It is hoped that the findings of this study may be beneficial for future studies of examinees' test-taking behaviors on multiple-choice test items. These behaviors have never been conclusively established or confirmed in psychometric research.

Table 1

Selected Values of x_is.

Assumption	x_1	x_2
1	1/2	1/4
2	1/4	1/8
3	0	0

Table 2

Selected Values of y_js

Assumption	y_1	y_2
1	1	1
2	1/2	2/3
3	1/4	1/3
4	0	0

Table 3.

Specification of the models.

Model[a]	x_1	x_2	y_1	y_2
M_{11}	1/2	1/4	1	1
M_{12}	1/2	1/4	1/2	2/3
M_{13}	1/2	1/4	1/4	1/3
M_{14}	1/2	1/4	0	0
M_{21}	1/4	1/8	1	1
M_{22}	1/4	1/8	1/2	2/3
M_{23}	1/4	1/8	1/4	1/3
M_{24}	1/4	1/8	0	0
M_{31}	0	0	1	1
M_{32}	0	0	1/2	2/3
M_{33}	0	0	1/4	1/3
M_{34}	0	0	0	0

Note. [a]M_{km} = the model incorporating assumption *K* about the x_is and assumption *m* about the y_js

Table 4

Spearman rank correlations between finite state scores and number-right scores

	M_{11}[a]	M_{12}	M_{13}	M_{14}	M_{21}	M_{22}
C	.996	.994	.994	.993	.996	.993

	M_{23}	M_{24}	M_{31}	M_{32}	M_{33}	M_{34}
C	.992	.992	.996	.991	.991	.990

Notes. N= 5274

C= Number-right scores expressed as proportions of correct answers

[a]M_{km} = finite state model incorporating assumption *k* about the x_is and assumption *m* about the y_js

Table 5

Descriptive Statistics of Scores

Models[a]	Mean	S.D	Median
M_{11}	0.31	0.16	0.31
M_{12}	0.31	0.16	0.31
M_{13}	0.31	0.16	0.31
M_{14}	0.31	0.16	0.31
M_{21}	0.37	0.17	0.39
M_{22}	0.37	0.17	0.39
M_{23}	0.37	0.17	0.39
M_{24}	0.37	0.17	0.39
M_{31}	0.43	0.17	0.46
M_{32}	0.44	0.17	0.46
M_{33}	0.44	0.17	0.46
M_{34}	0.44	0.17	0.46
C	0.54	0.16	0.54
RC	0.38	0.21	0.38

Notes. N= 5274

C= Number-right scores RC= Rescaled number-right scores

[a]M_{km} =the model incorporating assumption *k* about the x_is and assumption *m* about the y_js.

Table 6

Means and standard deviations of scores by gender

Gender		M_{11}[a]	M_{12}	M_{13}	M_{14}	M_{21}	M_{22}	M_{23}	M_{24}	M_{31}	M_{32}	M_{33}	M_{34}	RC[b]
male	M	.30	.31	.31	.31	.37	.37	.37	.37	.43	.44	.44	.44	.38
N=2744	SD	.15	.15	.15	.15	.16	.16	.16	.16	.16	.16	.16	.16	.19
female	M	.31	.32	.32	.32	.38	.38	.38	.38	.44	.44	.44	.44	.39
N=2506	SD	.17	.17	.17	.17	.17	.18	.19	.19	.19	.19	.19	.19	.22

Notes. [a]M_{km} = the model incorporating assumption k about the x_is and assumption m about the y_js [b]RC = rescaled number-right scores

Table 7

Means and standard deviations of scores by ethnic group

Race		M_{11}[a]	M_{12}	M_{13}	M_{14}	M_{21}	M_{22}	M_{23}	M_{24}	M_{31}	M_{32}	M_{33}	M_{34}	RC[b]
Asian	M	.33	.33	.33	.33	.39	.40	.40	.40	.46	.46	.46	.46	.41
n=201	SD	.16	.16	.16	.16	.17	.17	.17	.17	.16	.16	.16	.16	.20
Indian	M	.30	.30	.30	.30	.36	.36	.36	.36	.42	.43	.43	.43	.37
n=81	SD	.17	.17	.17	.17	.18	.18	.18	.18	.18	.18	.18	.18	.22
Black	M	.23	.23	.23	.23	.29	.29	.29	.29	.36	.36	.36	.36	.29
n=1487	SD	.14	.14	.14	.14	.15	.16	.16	.16	.16	.16	.16	.16	.18
Hispanic	M	.33	.34	.34	.34	.40	.40	.40	.40	.47	.47	.47	.47	.42
n=74	SD	.16	.16	.16	.16	.17	.17	.17	.17	.17	.17	.17	.17	.20
White	M	.34	.34	.34	.35	.41	.41	.41	.41	.47	.48	.48	.48	.43
n=3248	SD	.16	.16	.16	.16	.17	.17	.17	.17	.17	.17	.17	.17	.20

Notes. [a]M_{km} = the model incorporating assumption k about the x_is and assumption m about the y_js [b]RC = rescaled number-right scores

Table 8

Means and standard deviations of scores by GPA level

GPA		M_{11}[a]	M_{12}	M_{13}	M_{14}	M_{21}	M_{22}	M_{23}	M_{24}	M_{31}	M_{32}	M_{33}	M_{34}	RC[b]
A	M	.46	.47	.47	.47	.52	.53	.53	.53	.58	.58	.58	.58	.57
n=189	SD	.17	.18	.18	.18	.18	.18	.18	.18	.19	.18	.18	.18	.21
B	M	.38	.38	.38	.39	.45	.45	.46	.46	.51	.52	.52	.52	.48
n=954	SD	.14	.14	.14	.14	.15	.15	.15	.15	.14	.14	.14	.14	.18
C+	M	.28	.28	.28	.28	.34	.34	.34	.34	.41	.41	.41	.41	.34
n=1460	SD	.14	.14	.14	.14	.16	.16	.16	.16	.16	.16	.16	.16	.18
C-	M	.23	.23	.23	.23	.28	.29	.29	.29	.35	.35	.35	.35	.28
n=893	SD	.14	.14	.14	.14	.16	.16	.17	.17	.17	.17	.17	.17	.18

Note. [a]M_{km} = the model incorporating assumption k about the x_is and assumption m about the y_js [b]RC= rescaled number-right scores

Table 9

Means and standard deviations of scores by years of science course work

Group		M_{11}[a]	M_{12}	M_{13}	M_{14}	M_{21}	M_{22}	M_{23}	M_{24}	M_{31}	M_{32}	M_{33}	M_{34}	RC[b]
3yrs+	M	.36	.37	.37	.37	.43	.43	.43	.44	.49	.50	.50	.50	.45
n=168	SD	.17	.17	.17	.17	.18	.18	.18	.18	.18	.18	.18	.18	.21
3yrs	M	.34	.35	.35	.35	.41	.42	.42	.42	.48	.48	.48	.48	.43
n=1578	SD	.15	.15	.15	.15	.16	.16	.16	.16	.16	.16	.16	.16	.19
2yrs+	M	.26	.26	.26	.26	.32	.32	.33	.33	.39	.39	.39	.39	.33
n=1706	SD	.15	.15	.15	.15	.16	.17	.17	.17	.17	.17	.17	.17	.19
2yrs-	M	.22	.22	.22	.22	.27	.27	.27	.27	.33	.34	.34	.34	.27
n=158	SD	.16	.16	.16	.16	.18	.18	.18	.18	.19	.19	.19	.19	.20

Note. [a]M_{km} = the model incorporating assumption k about x_i and assumption m about y_j [b]RC = rescaled number-right scores

Table 10

Means and standard deviations of scores by omissiveness level

Group		M_{11}[a]	M_{12}	M_{13}	M_{14}	M_{21}	M_{22}	M_{23}	M_{24}	M_{31}	M_{32}	M_{33}	M_{34}	RC[b]
least	M	.23	.23	.23	.23	.29	.29	.29	.29	.36	.36	.36	.36	.30
n=2588	SD	.13	.13	.13	.13	.15	.15	.15	.15	.16	.16	.16	.16	.17
moderate	M	.38	.38	.39	.39	.45	.45	.46	.46	.51	.52	.52	.52	.47
n=2078	SD	.16	.16	.16	.16	.16	.16	.16	.16	.16	.16	.16	.16	.20
most	M	.37	.38	.39	.39	.44	.46	.46	.46	.51	.53	.53	.53	.41
n=253	SD	.15	.16	.16	.16	.16	.16	.16	.15	.15	.14	.14	.14	.21

Note. [a]M_{km} = the model incorporating assumption k about the x_is and assumption m about the y_js [b]RC = rescaled number-right scores

Table 11

Summary table for repeated measures analysis of variance on type of scoring and gender

Source	SS	df	MS	F
Between groups				
groups	0.99	1	0.99	2.53
error	2051.23	5248	0.39	
Within group				
scores	179.42	12	14.95	45611.07*
scores x group	0.15	12	0.01	38.50*
error (score)	20.64	62976	0.0003	

Note. * $p<.0001$

Table 12

Summary table for repeated measures analysis of variance on type of scoring and ethnic group

Source	SS	df	MS	F
Between groups				
groups	185.68	4	46.42	131.36*
error	1797.36	5086	0.35	
Within group				
scores	27.11	12	2.26	7142.40*
scores x group	0.71	48	0.01	46.57*
error (score)	19.31	61032	0.0003	

Note. * p<.0001

Table 13

Summary table for repeated measures analysis of variance on type of scoring and GPA level

Source	SS	df	MS	F
Between groups				
groups	239.62	3	79.87	252.36*
error	1105.27	3492	0.32	
Within group				
scores	64.85	12	5.40	17366.56*
scores x group	1.01	36	0.03	90.39*
error (scores)	13.04	41904	0.0003	

Note. * p<.0001

Table 14

Summary table for repeated measures analysis of variance on type of scoring and years of science course work

Source	SS	df	MS	F
Between groups				
groups	113.21	3	37.74	106.43*
error	1278.62	3606	0.35	
Within group				
scores	38.43	12	3.20	9807.34*
scores x group	0.46	36	0.01	38.85*
error (scores)	14.13	43272	0.0003	

Note. * p<.0001

Table 15

Summary table for repeated measures analysis of variance on type of scoring and omissiveness level

Source	SS	df	MS	F
Between groups				
groups	392.80	2	196.40	629.44*
error	1533.92	4916	0.31	
Within group				
scores	68.07	12	5.67	18495.34*
scores x group	1.59	24	0.07	215.65*
error (scores)	18.09	58992	0.0003	

Note. * p<.0001

Table 16

Descriptive Statistics for γ_0

Model[a]	Median	Q[b]
M_{11}	0.85	0.31
M_{21}	0.78	0.51
M_{31}	0.65	0.83
M_{12}	0.93	0.11
M_{22}	0.92	0.12
M_{32}	0.92	0.13
M_{13}	0.95	0.08
M_{23}	0.95	0.07
M_{33}	0.95	0.07
M_{14}	0.96	0.06
M_{24}	0.96	0.06
M_{34}	0.96	0.05

Notes. N=5274

[a] M_{km} = the model incorporating assumption *k* about the x_is and assumption *m* about the y_js [b] Q = interquartile range

Table 17

Means and standard deviations of γ_0 by gender

Gender		M_{11}[a]	M_{12}	M_{13}	M_{14}	M_{21}	M_{22}	M_{23}	M_{24}	M_{31}	M_{32}	M_{33}	M_{34}
male	M	.74	.87	.90	.92	.63	.86	.90	.93	.52	.85	.90	.93
n=2744	SD	.27	.14	.10	.08	.34	.16	.11	.08	.37	.17	.11	.08
female	M	.73	.88	.91	.93	.63	.87	.91	.93	.53	.86	.91	.93
n=2506	SD	.30	.14	.10	.08	.35	.16	.10	.08	.38	.17	.10	.08

Note. [a]M_{km} = the model incorporating assumption K about the x_is and assumption m about the y_js

Table 18

Means and standard deviations of γ_0 by ethnic group

Race		M_{11}[a]	M_{21}	M_{31}	M_{12}	M_{22}	M_{32}	M_{13}	M_{23}	M_{33}	M_{14}	M_{24}	M_{34}
Asian	M	.71	.59	.47	.86	.85	.84	.90	.90	.90	.92	.92	.92
n=201	SD	.29	.34	.37	.14	.16	.17	.11	.11	.11	.09	.08	.08
Indian	M	.74	.64	.54	.90	.90	.90	.92	.92	.92	.94	.94	.94
n=81	SD	.28	.36	.39	.10	.11	.11	.07	.07	.07	.06	.06	.05
Black	M	.83	.76	.67	.90	.89	.88	.92	.92	.92	.94	.94	.94
n=1487	SD	.22	.28	.33	.12	.13	.14	.09	.09	.09	.08	.07	.07
Hispanic	M	.71	.58	.47	.87	.85	.84	.90	.90	.90	.92	.93	.93
n=74	SD	.28	.36	.38	.16	.17	.18	.11	.11	.11	.08	.08	.08
White	M	.69	.58	.46	.86	.85	.84	.90	.90	.90	.92	.92	.93
n=3248	SD	.30	.35	.37	.15	.17	.18	.11	.11	.11	.08	.08	.08

Notes. [a]M_{km} = the model incorporating assumption k about the x_is and assumption m about the y_js

Table 19

Means and standard deviations of γ_0 by GPA level

GPA		M_{11}[a]	M_{21}	M_{31}	M_{12}	M_{22}	M_{32}	M_{13}	M_{23}	M_{33}	M_{14}	M_{24}	M_{34}
A	M	.49	.33	.23	.80	.79	.78	.87	.87	.87	.90	.91	.91
n=189	SD	.34	.36	.33	.20	.21	.22	.13	.13	.13	.10	.09	.09
B	M	.64	.51	.38	.84	.82	.81	.88	.88	.88	.91	.91	.92
n=954	SD	.32	.35	.35	.17	.19	.20	.12	.13	.13	.10	.09	.09
C+	M	.78	.69	.58	.89	.87	.86	.91	.91	.91	.93	.93	.93
n=1460	SD	.25	.31	.36	.13	.15	.16	.10	.10	.10	.08	.08	.08
C-	M	.83	.77	.69	.91	.90	.90	.93	.93	.93	.94	.94	.95
n=893	SD	.21	.28	.33	.11	.12	.13	.08	.08	.09	.07	.07	.06

Note. [a]M_{km} = the model incorporating assumption k about x_i and assumption m about y_j

Table 20

Means and standard deviations of γ_0 by years of science course work

Group		M_{11}[a]	M_{21}	M_{31}	M_{12}	M_{22}	M_{32}	M_{13}	M_{23}	M_{33}	M_{14}	M_{24}	M_{34}
3yrs+	M	.62	.51	.40	.82	.81	.79	.87	.87	.87	.90	.90	.91
n=168	SD	.35	.38	.38	.19	.21	.22	.13	.14	.14	.11	.10	.10
3yrs	M	.69	.58	.46	.86	.85	.84	.90	.90	.90	.92	.92	.93
n=1578	SD	.30	.35	.37	.16	.17	.18	.11	.11	.11	.09	.08	.08
2yrs+	M	.80	.71	.61	.89	.88	.87	.92	.92	.92	.93	.93	.94
n=1706	SD	.24	.32	.36	.12	.15	.16	.10	.10	.10	.08	.08	.08
2yrs-	M	.80	.72	.63	.89	.88	.87	.91	.91	.91	.93	.93	.93
n=158	SD	.24	.32	.38	.12	.14	.15	.10	.10	.10	.08	.08	.07

Note. [a]M_{km} = the model incorporating assumption k about the x_is and assumption m about the y_js

Table 21

Means and standard deviations of γ_0 by omissiveness level

Group		M_{11}[a]	M_{21}	M_{31}	M_{12}	M_{22}	M_{32}	M_{13}	M_{23}	M_{33}	M_{14}	M_{24}	M_{34}
least	M	.93	.90	.82	.96	.95	.95	.97	.97	.97	.97	.97	.98
n=2588	SD	.05	.09	.14	.01	.02	.02	.01	.01	.02	.01	.01	.01
moderate	M	.57	.39	.23	.83	.81	.80	.87	.87	.87	.90	.90	.91
n=2078	SD	.27	.31	.30	.08	.09	.10	.06	.06	.06	.04	.05	.05
most	M	.11	.05	.01	.40	.34	.30	.56	.55	.54	.65	.66	.67
n=253	SD	.18	.12	.08	.17	.17	.17	.11	.12	.12	.09	.08	.08

Note. [a]M_{km} = the model incorporating assumption k about the x_is and assumption m about the y_js

Table 22

Summary table for repeated measures analysis of variance on γ_0 from 12 finite state models and gender

Source	SS	df	MS	F
Between groups				
groups	0.28	1	0.28	0.97
error	1533.25	5248	0.29	
Within group				
models	1008.37	11	91.67	5816.73*
models x group	0.32	11	0.03	1.86
error (score)	909.78	57728	0.01	

Note. * $p<.0001$

Table 23

Summary table for repeated measures analysis of variance on γ_0 from 12 finite state models and ethnic group

Source	SS	df	MS	F
Between groups				
groups	49.46	4	12.36	43.86*
error	1433.89	5086	0.28	
Within group				
models	154.13	11	14.01	950.95*
models x group	53.86	44	1.22	83.08*
error (model)	824.37	55946	0.01	

Note. * $p<.0001$

Table 24

Summary table for repeated measures analysis of variance on γ_0 from 12 finite state models and GPA level

Source	SS	df	MS	F
Between groups				
groups	86.79	3	28.93	103.75*
error	973.75	3492	0.28	
Within group				
models	497.94	11	45.27	3379.5*
models x group	79.90	33	2.42	180.76*
error (models)	514.52	38412	0.013	

Note. * $p<.0001$

Table 25

Summary table for repeated measures analysis of variance on γ_0 from 12 finite state models and years of science course work

Source	SS	df	MS	F
Between groups				
groups	30.65	3	10.21	34.55*
error	1066.12	3606	0.29	
Within group				
models	222.59	11	20.23	1375.95*
models x group	28.59	33	0.87	58.92*
error (models)	583.35	39666	0.015	

Note. * p<.0001

Table 26

Summary table for repeated measures analysis of variance on γ_0 from 12 finite state models and omissiveness level

Source	SS	df	MS	F
Between groups				
groups	1140.29	2	570.14	8313.10*
error	337.16	4916	0.07	
Within group				
models	567.75	11	51.61	6777.75*
models x group	466.15	22	21.19	2782.43*
error (models)	411.80	54076	0.008	

Note. * p<.0001

Table 27
Correlations between number of embedded omissions, omissiveness and γ_0

Model[a]	# of Embs[b]	Omissiveness
M_{11}	-.53	-.69
M_{21}	-.49	-.61
M_{31}	-.46	-.54
M_{12}	-.79	-.94
M_{22}	-.81	-.94
M_{32}	-.83	-.93
M_{13}	-.87	-.96
M_{23}	-.88	-.96
M_{33}	-.88	-.96
M_{14}	-.92	-.96
M_{24}	-.92	-.96
M_{34}	-.91	-.96

Notes. N=5274

[a]M_{km}= the model incorporating assumption *k* about the x_is and assumption *m* about the y_j s [b]# of Embs = number of embedded omissions

Table 28
Correlations between finite state scores, number-right scores, and γ_0

Models[a]	Finite State scores	number-right scores
M_{11}	-.66	-.60
M_{21}	-.71	-.67
M_{31}	-.74	-.72
M_{12}	-.39	-.30
M_{22}	-.38	-.29
M_{32}	-.37	-.27
M_{13}	-.30	-.20
M_{23}	-.30	-.19
M_{33}	-.30	-.18
M_{14}	-.24	-.14
M_{24}	-.25	-.13
M_{34}	-.26	-.14

Note. the model incorporating assumption *k* about the x_is and assumption *m* about the y_js

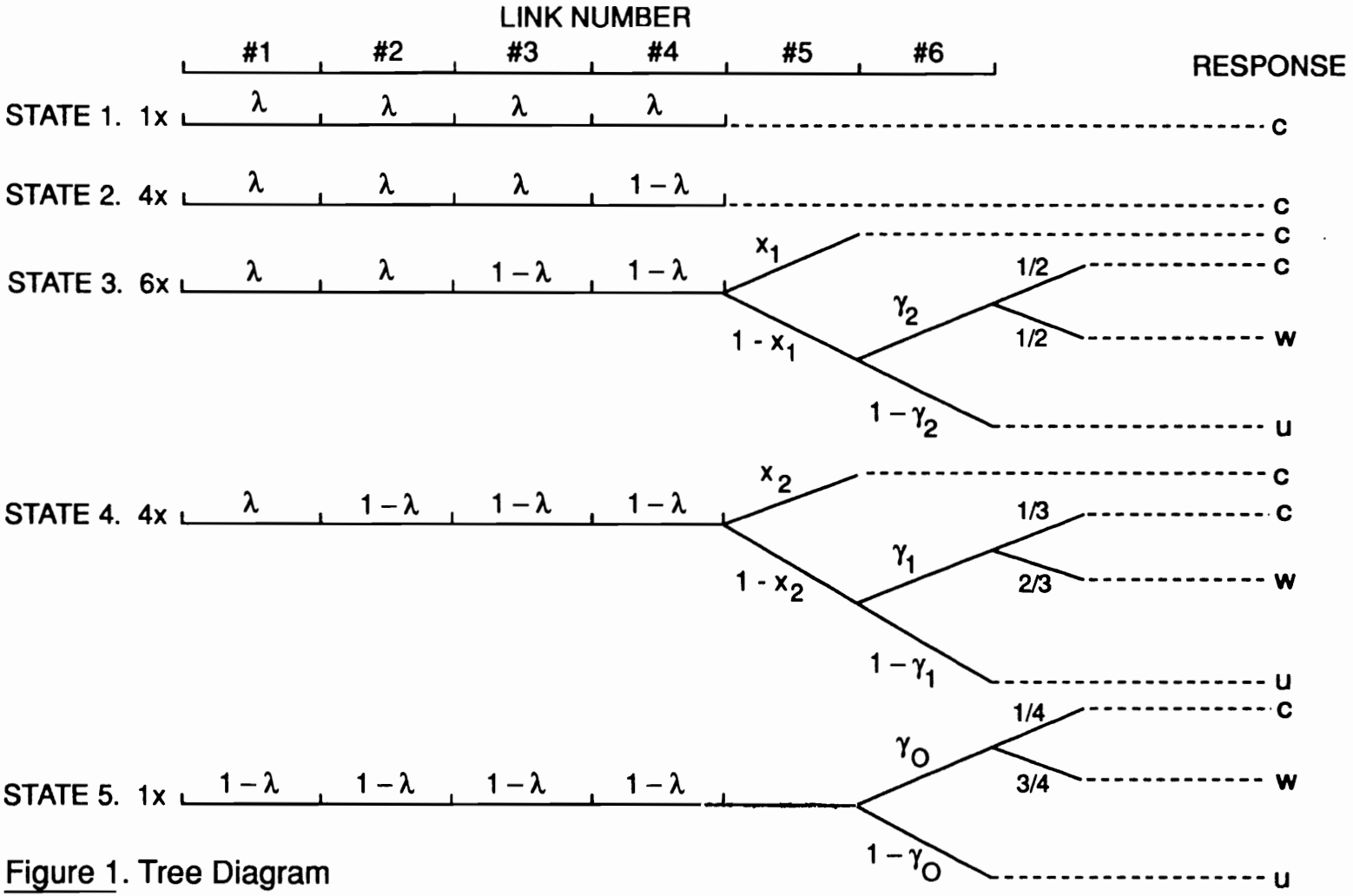

Figure 1. Tree Diagram

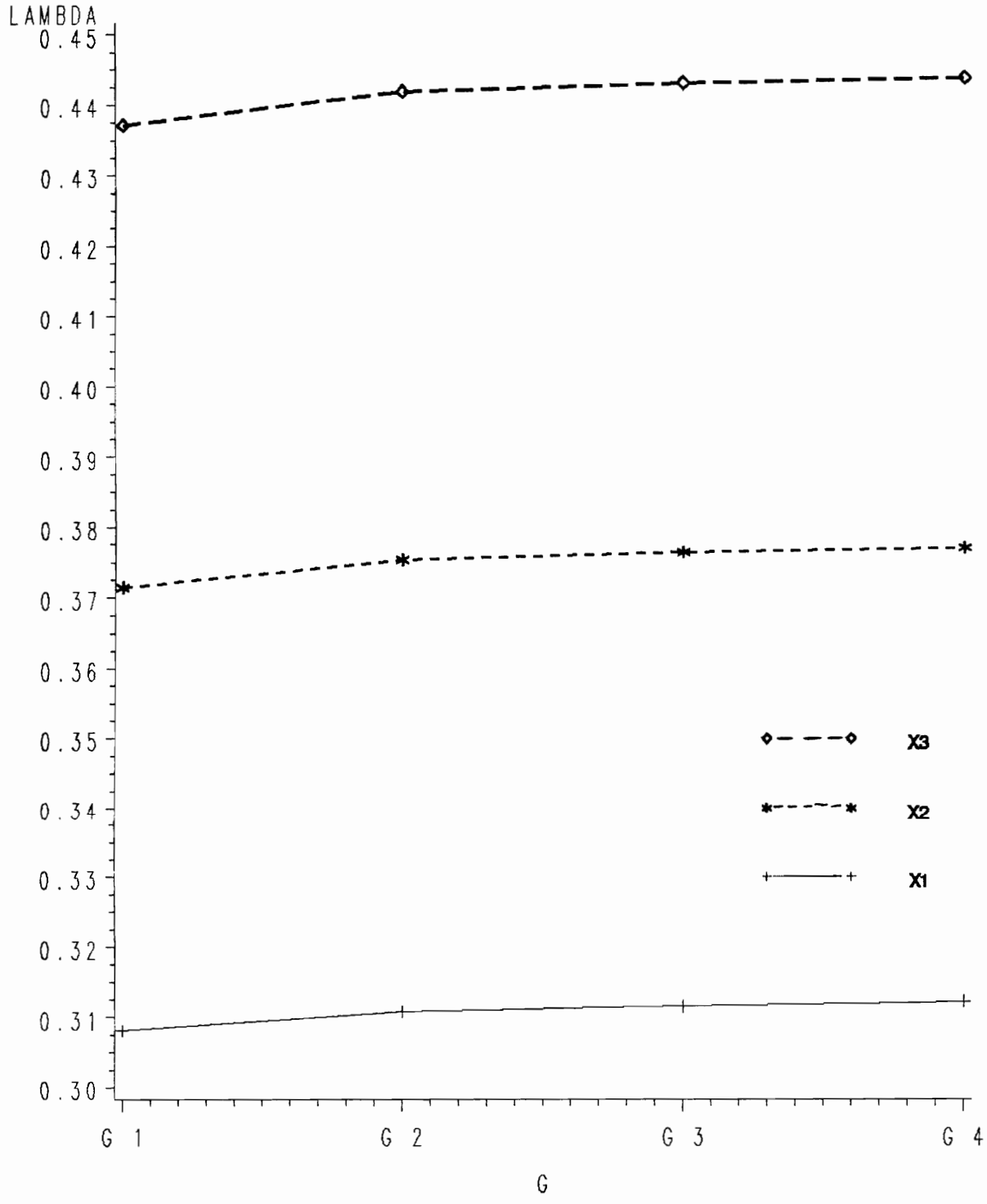

Figure 2. Mean scores from 12 finite state models

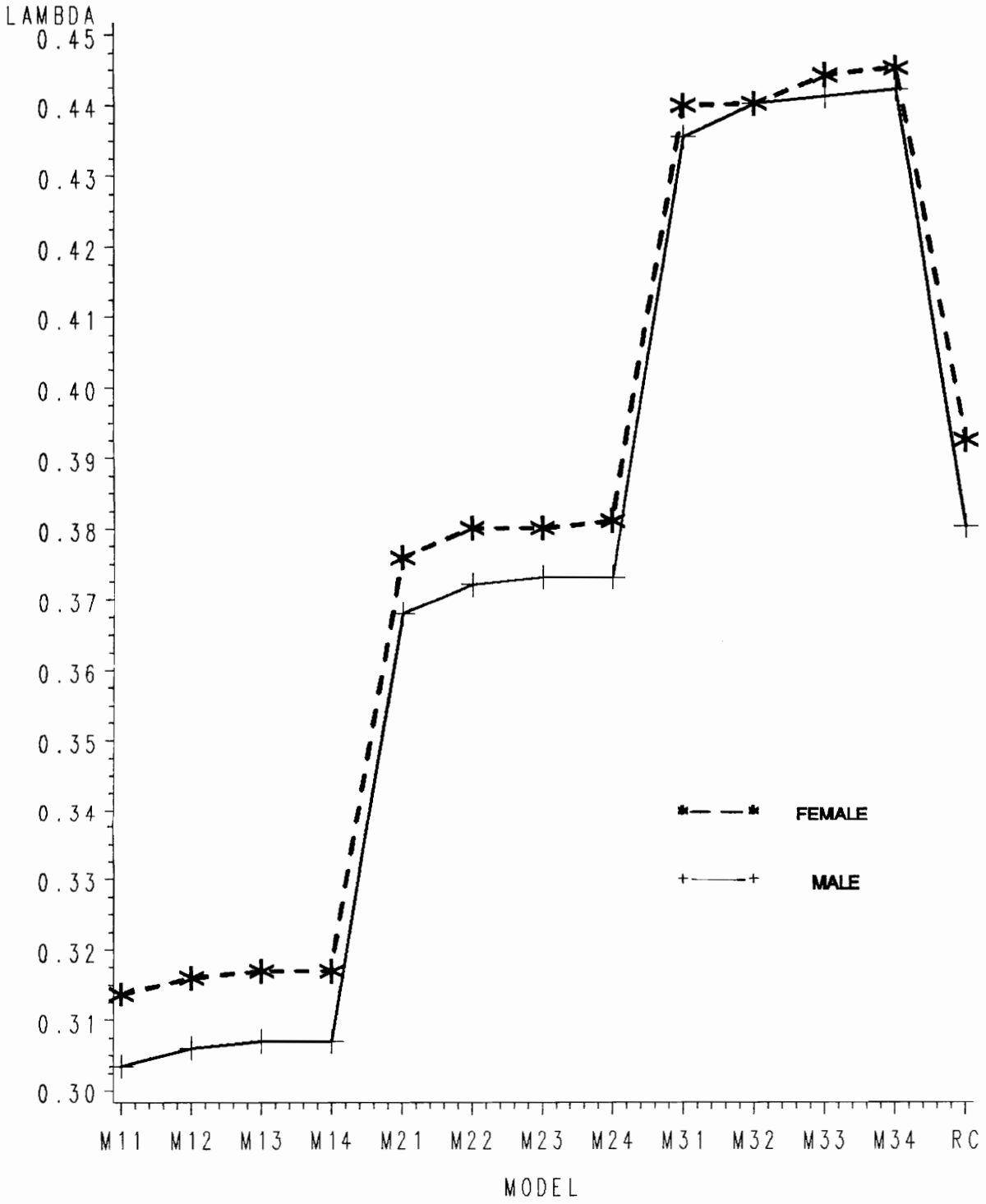

FIGURE 3. Mean scores by gender

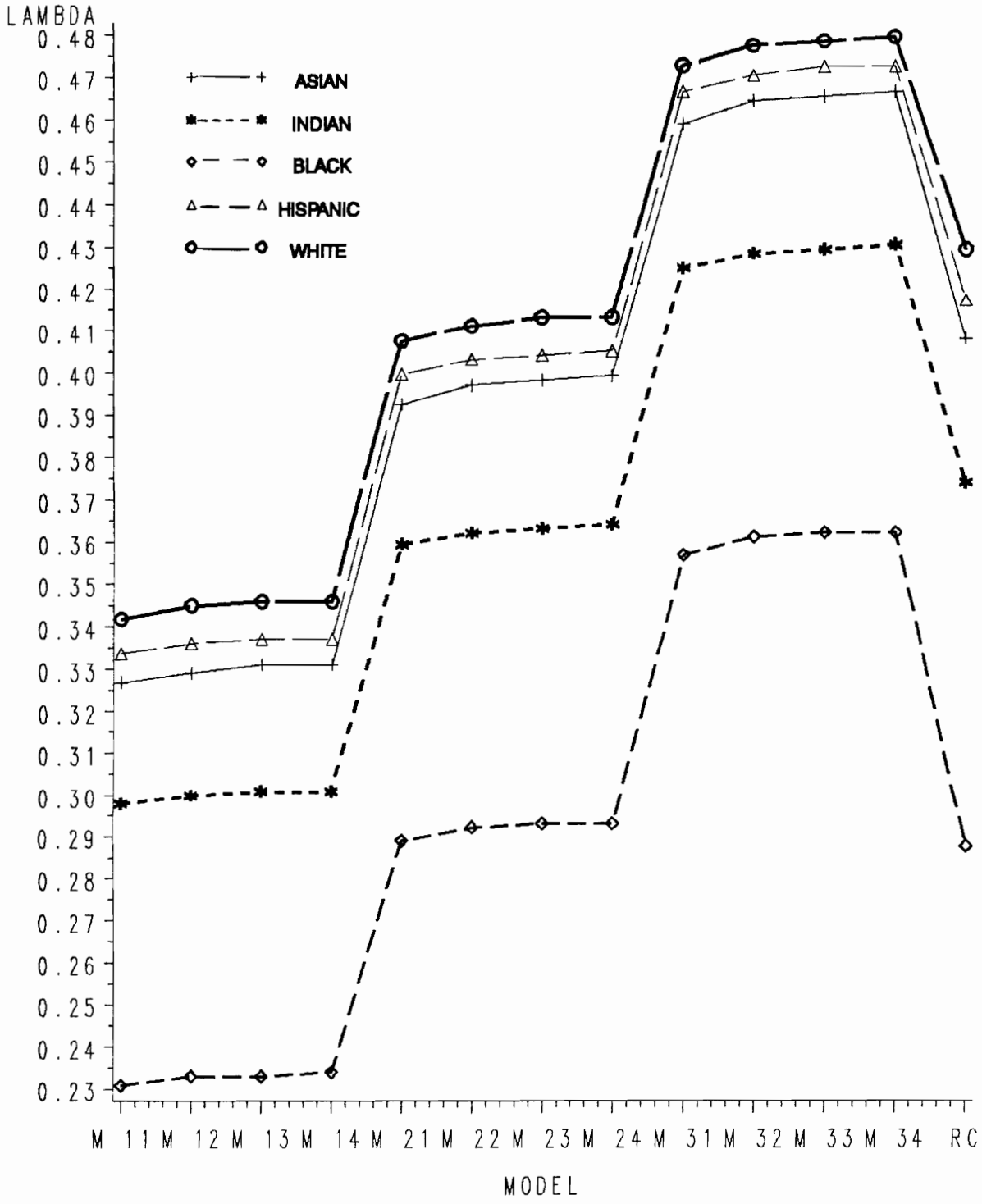

Figure 4. Mean scores by ethnic group

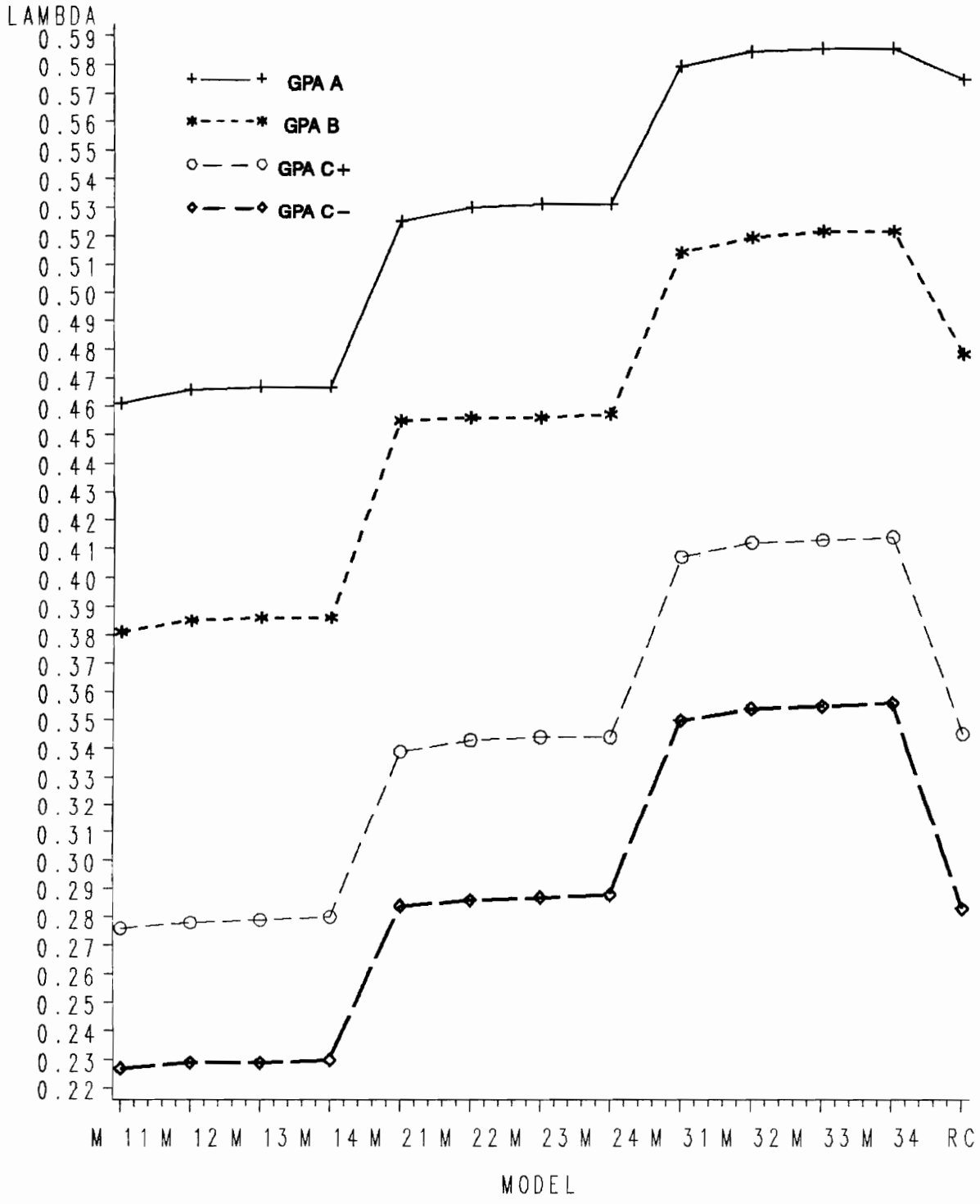

Figure 5. Mean scores by GPA level

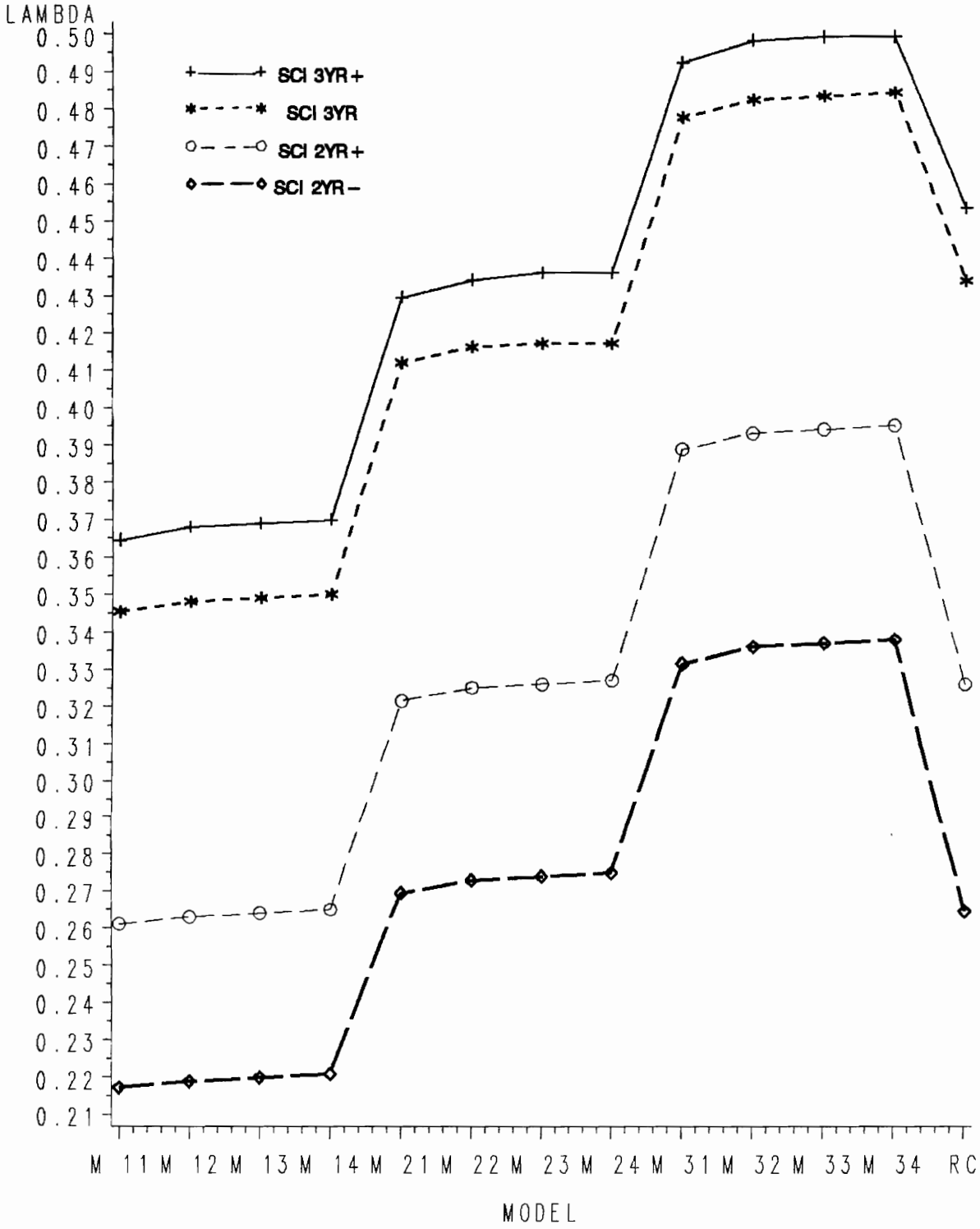

Figure 6. Mean scores by years of science course work

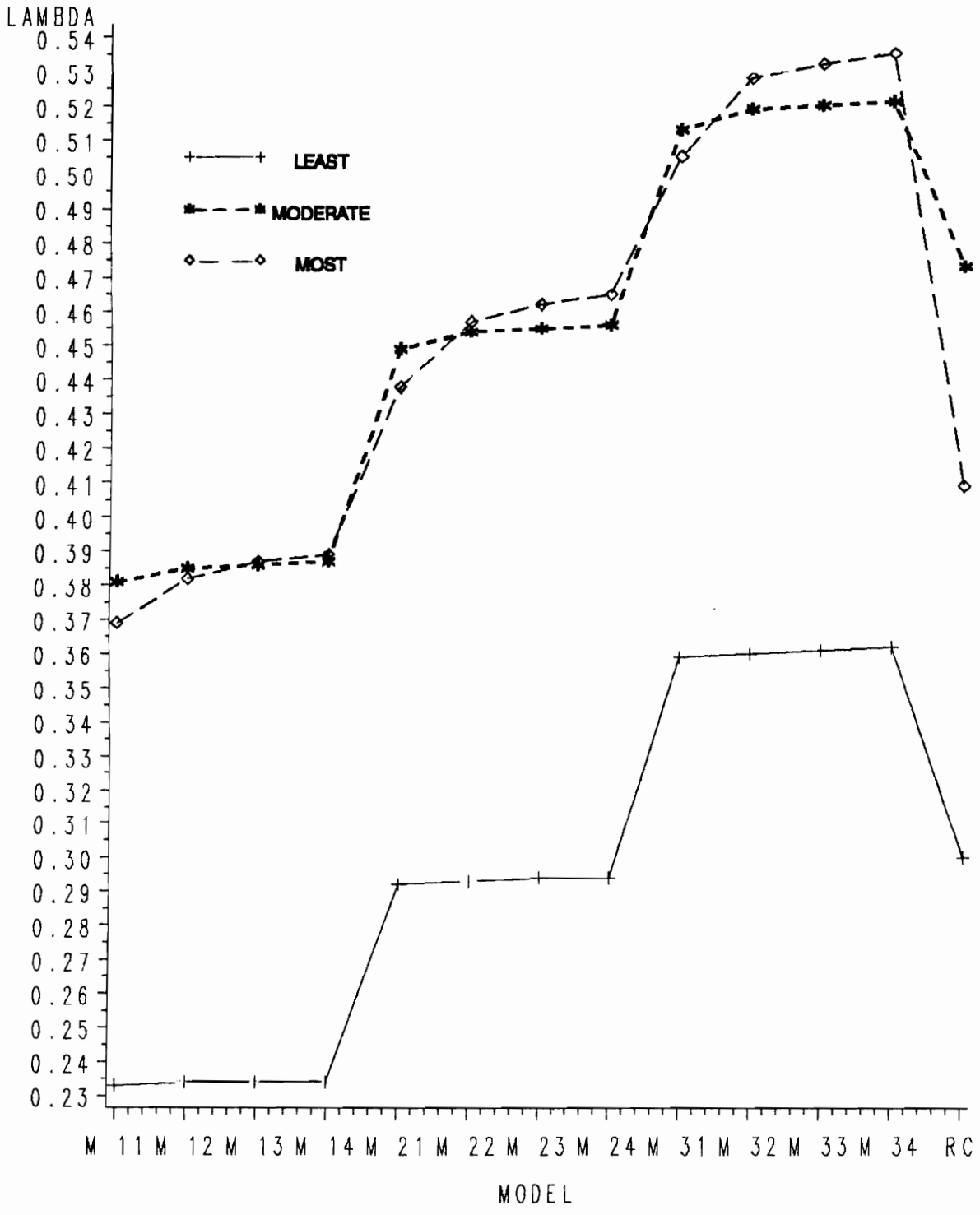

Figure 7. Mean scores by omissiveness level

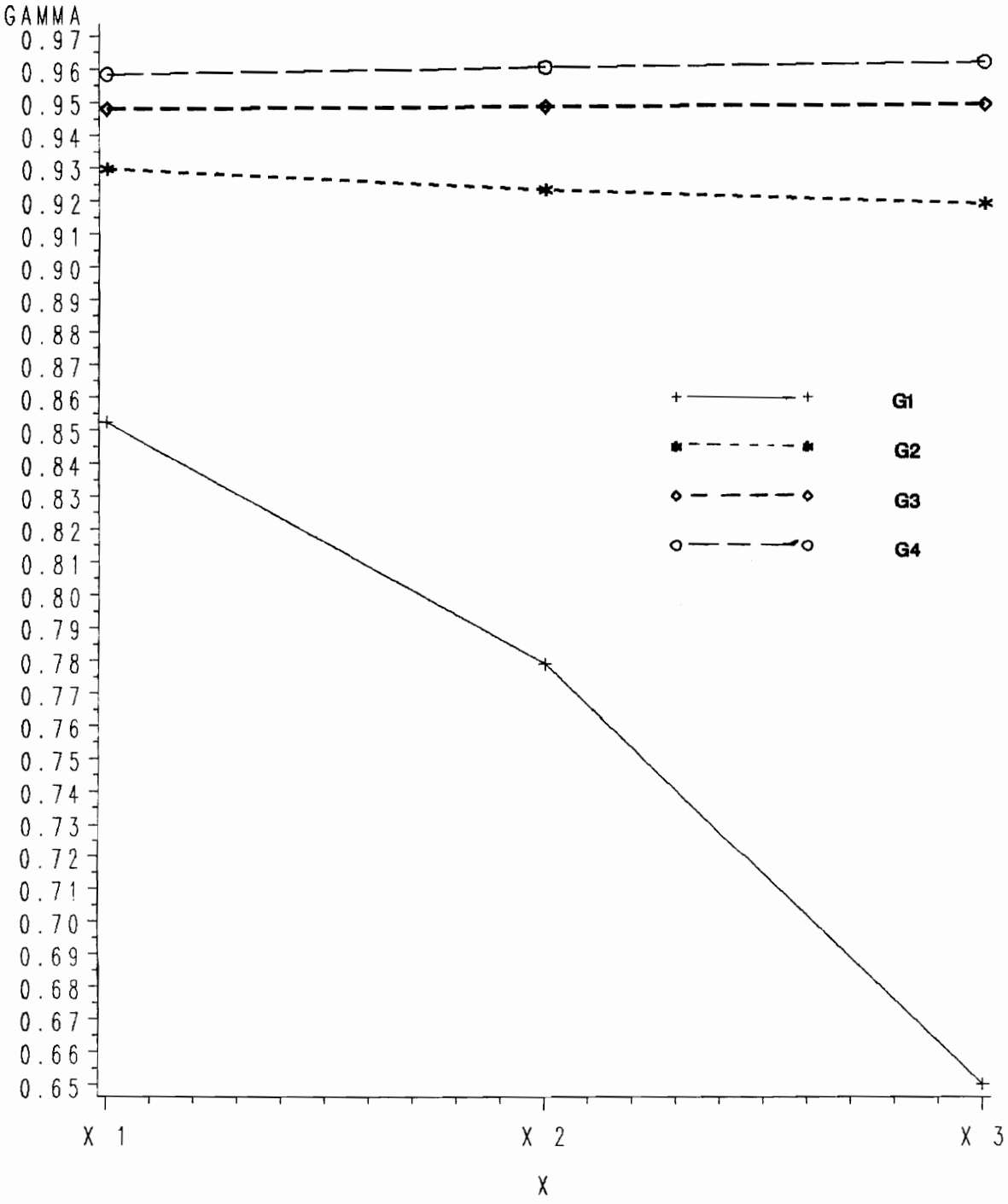

FIGURE 8. Medians of estimated willingness to guess from 12 finite state models

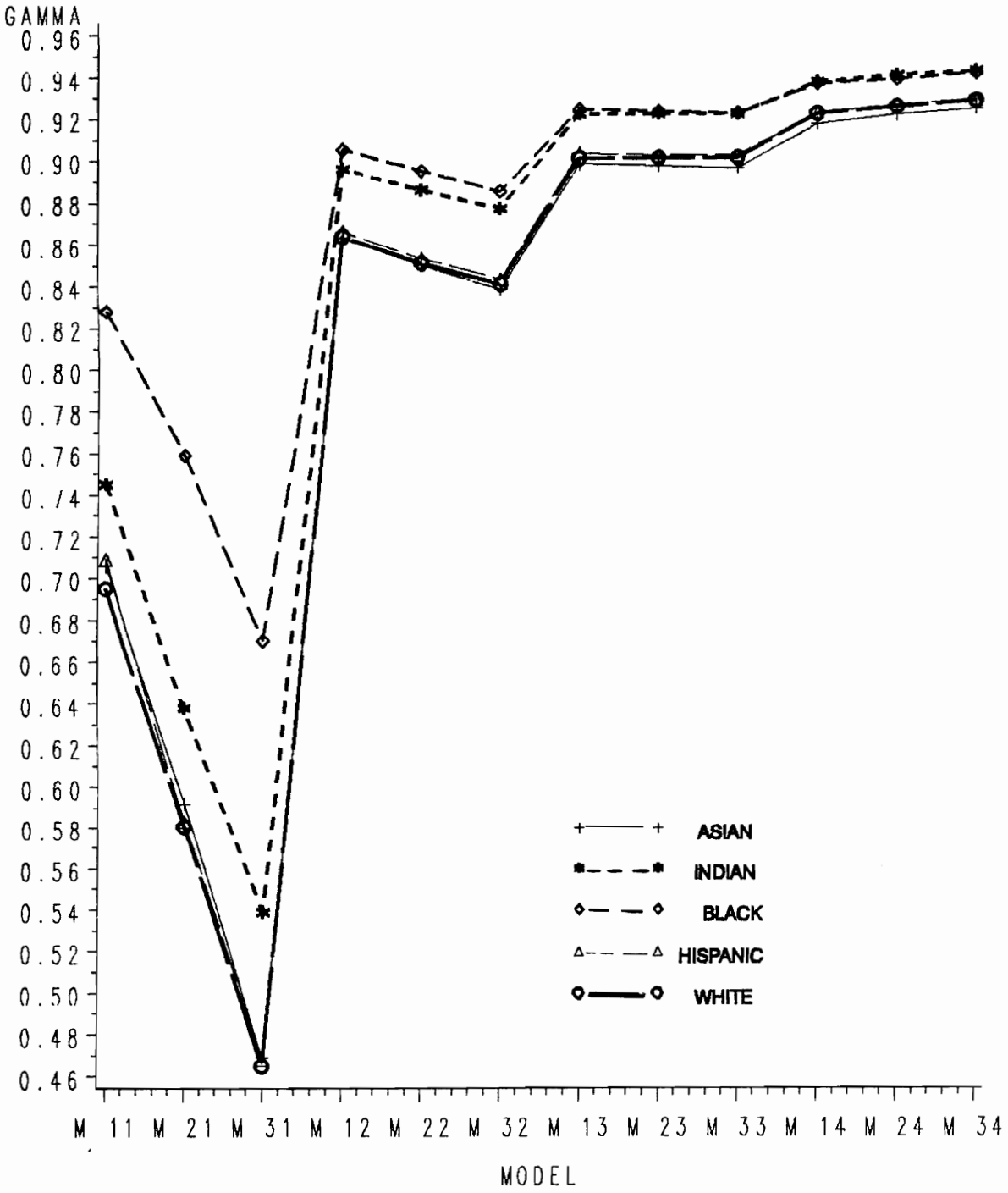

Figure 9. Means of estimated willingness to guess by ethnic group

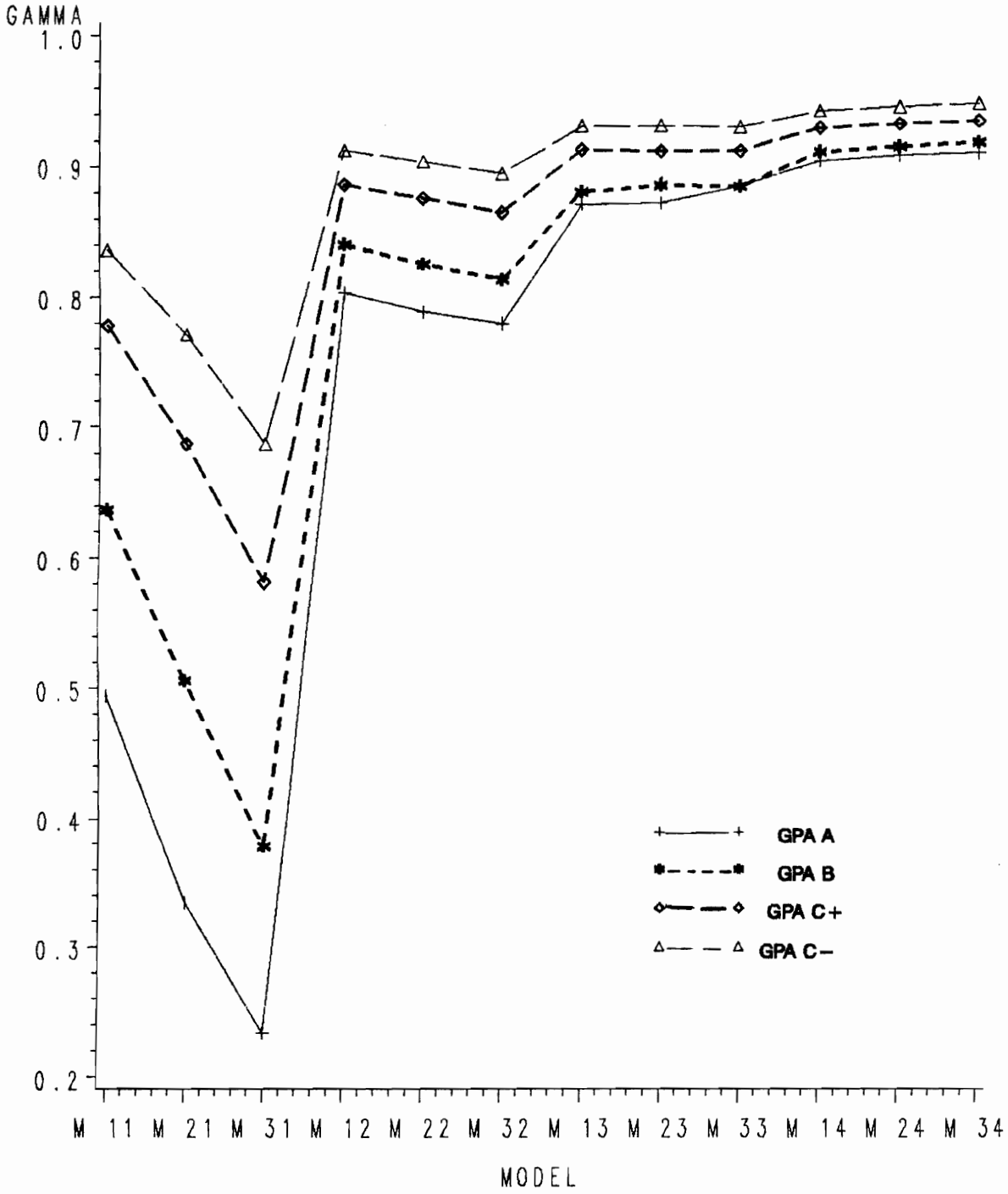

FIGURE 10. Means of estimated willingness to guess by GPA level

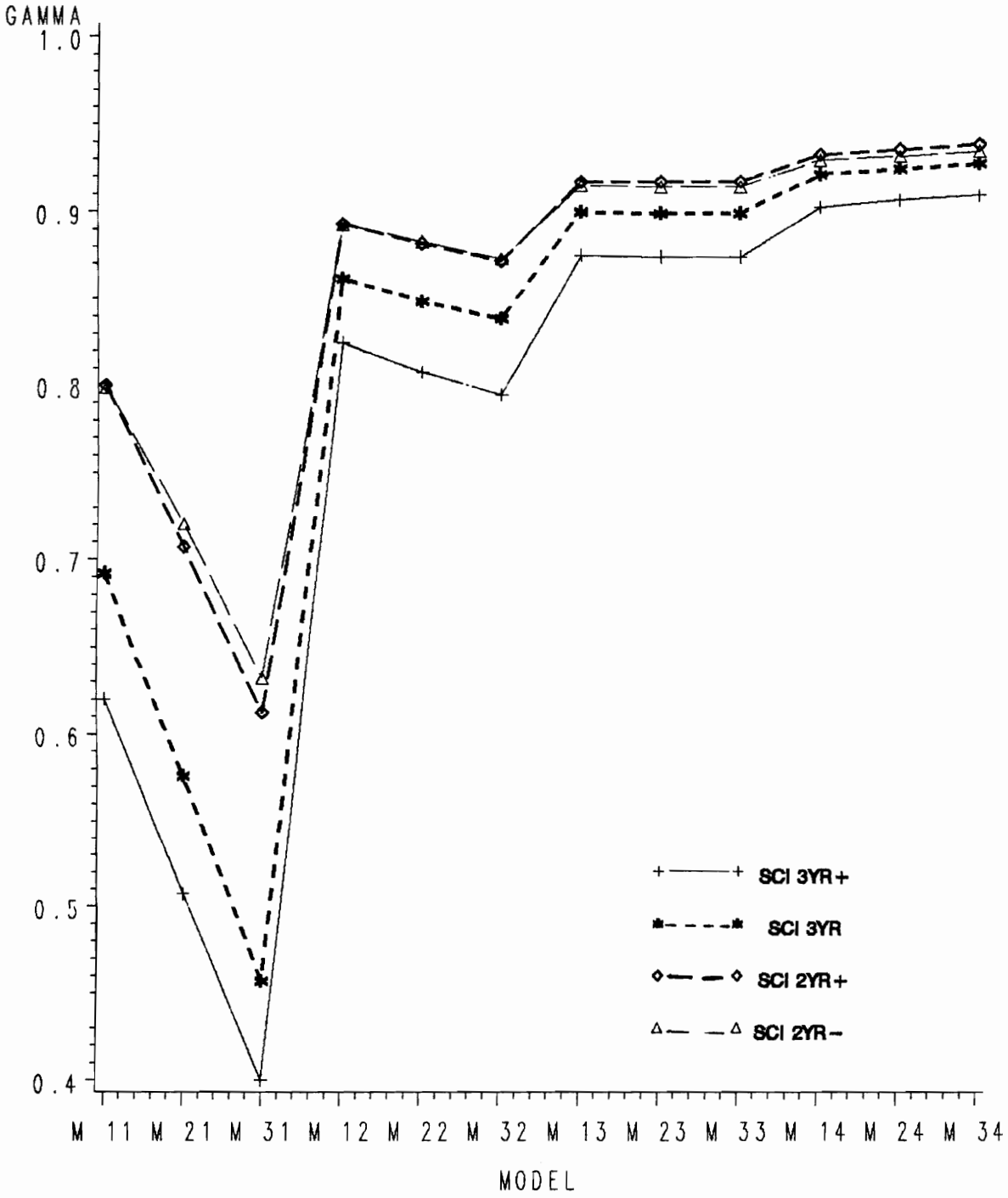

Figure 11. Means of estimated willingness to guess by years of science course work

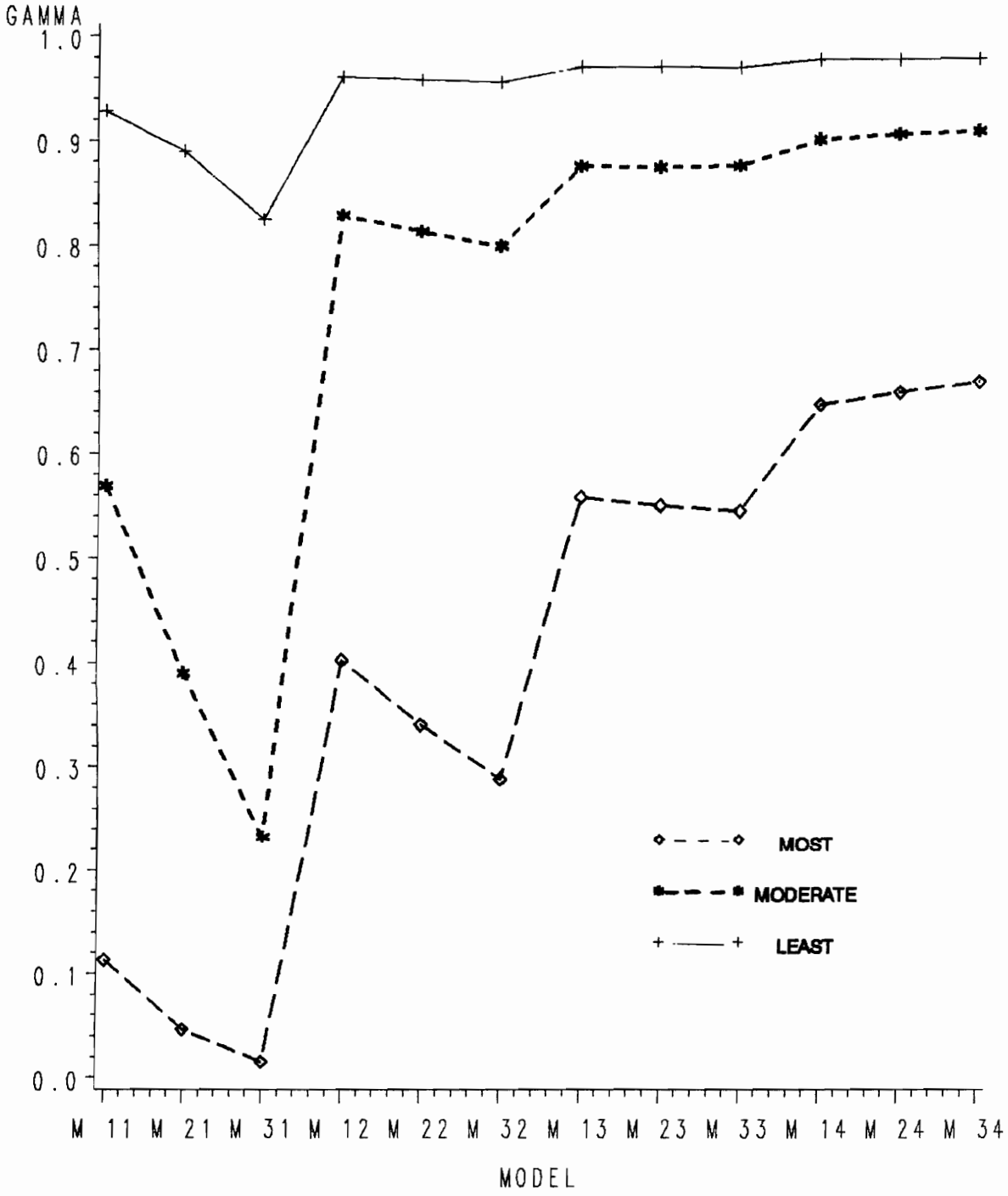

Figure 12. Means of estimated willingness to guess by omissiveness level

References

Bliss, L. B. (1980). A test of Lord's assumption regarding examinee guessing behavior on multiple choice tests using elementary school students. Journal of Educational Measurement, 17(2), 147-153.

Bruno, J. E. (1986). Assessing the knowledge base of students: An information theoretic approach of testing. Measurement and Evaluation in Counseling and Development, 19(3), 116-130.

Cheney, W. & Kincaid, D. (1985). Numerical Mathematics and Computing (2nd ed.). Pacific Grove, CA: Brooks/Cole.

Claudy, J. G. (1978). Biserial Weights: A new approach to test item option weighting. Applied Psychological Measurement, 2(1), 25-30.

Cross, L. H. & Frary, R. B. (1977). An empirical test of Lord's theoretical results regarding formula scoring of multiple choice tests. Journal of Educational Measurement, 14(4), 313-321.

Fisher, F. E. (1988). Effects of instruction for guessing on multiple-choice test performance. Educational Research Quarterly, 12(1), 6-9.

Frary, R. B. (1989). Partial-credit scoring methods for multiple-choice tests. Applied Measurement in Education, 2(1), 79-96.

García-Pérez, M. A. (1985 September). A finite state theory of performance in multiple-choice tests. Proceedings of the 16th European Psychological Group Meeting (pp. 55-67). Montpellier, France.

García-Pérez, M. A. (1987). A finite state theory of performance in multiple-choice tests. In E. E. Roskam & R. Suck (Eds.), Progress in mathematical psychology-I (pp. 455-462). Amsterdam: Elsevier.

García-Pérez, M. A. (1990). A comparison of two models of performance in objective tests: Finite state versus continuous distributions. British Journal of Mathematical and Statistical Psychology, 43, 73-91.

García-Pérez, M. A. (1991). In defense of "none of the above". Manuscript submitted for publication.

García-Pérez, M. A. & Frary, R. B. (1989). Psychometric properties of finite-state scores versus number-correct and formula scores: A simulation study. Applied Psychological Measurement, 13(4), 403-417.

García-Pérez, M. A. & Frary, R. B. (1991a). Testing finite models of performance in multiple-choice tests using items with 'none of the above' as an option. In J.-C.Falmagne & J.-P. Doignon (Eds.), Mathematical Psychology: Current developments. (pp. 273-291). New York: Springer-Verlag.

García-Pérez, M. A. & Frary, R. B. (1991b). Finite state polynomic item characteristic curves. British Journal of Mathematical and Statistical Psychology, 44, 45-73.

Ghiselli, E. E., Campbell, J. P., & Zedack, S. (1981). Measurement theory for the behavioral sciences. New York: Freeman.

Haladyna, T. M. & Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. Applied Measurement in Education, 2(1), 37-50.

Hutchinson, T. P. (1985 July). Predicting performance in variants of multiple choice test. Paper presented at the fourth annual meeting of the Psychometric Society and the Classification Societies. Cambridge, England .

Jaradat, D. & Sawaged, S. (1986). The subset selection techniques for multiple choice tests. An empirical inquiry. Journal of Educational Measurement, 23(4), 369-376.

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Loyd, B. H. (1988). Implications of item response theory for the measurement practitioner. Applied Measurement in Education, 1, 135-143.

Marso, R. N. (1985 October). Testing practice and test item preferences of classroom teachers. Paper presented at the annual meeting of the Mid Western Educational Research Association , Chicago, IL .

Ott, L. (1988). An introduction to statistical methods and data analysis (3rd ed.). Boston, MA: PWS-KENT.

SAS Institute, Inc. (1989). SAS/STAT user's guide. Cary, NC: author.

Scannell, D. P. (Ed.) (1986). Tests of achievement and proficiency. Chicago: Riverside Publishing Company.

Suen, H. K. (1990). Principles of test theories. Hillsdale, NJ: LEA

Worthen, B. R. & Sanders, J. R. (1987). Educational evaluation: Alternative approaches and practical guidelines. White plains, NY: Longman.

Wilcox, R. R. & Wilcox, K. T. (1988). Models of decision making process for multiple-choice test items: An analysis of spatial ability. Journal of Educational Measurement, 25(2), 125-136.

Appendix A

Equation 1.

$$c = \lambda^4 + 4\lambda^3 (1-\lambda) + 6\lambda^2 (1-\lambda)^2 x_1 + 6\lambda^2 (1-\lambda)^2 (1-x_1) \gamma_2 /2 + 4\lambda (1-\lambda)^3 x_2 + 4\lambda (1-\lambda)^3 (1-x_2)\gamma_1 /3 + (1-\lambda)^4 \gamma_o /4$$

Equation 2.

$$w = 6\lambda^2 (1-\lambda)^2 (1-x_1) \gamma_2 /2 + 4\lambda (1-\lambda)^3 (1-x_2) 2/3 \gamma_1 + (1-\lambda)^4 3/4 \gamma_o$$

Equation 3.

$$u = 6\lambda^2 (1-\lambda)^2 (1-x_1) (1-\gamma_2) + 4\lambda(1-\lambda)^3 (1-x_2) (1-\gamma_1) + + (1-\lambda)^4 (1-\gamma_o)$$

Equation 4. Finite state scoring polynomial.

$$\begin{aligned} 0 = {} & \lambda^4 (-3.+6x_1+3y_2-3x_1y_2-4x_2-4/3.y_1+4/3.x_2y_1) \\ & +\lambda^3 (4.-12x_1-6y_2+6x_1y_2+12x_2+4y_1-4x_2y_1) \\ & +\lambda^2 (6x_1+3y_2-3x_1y_2-12x_2-4y_1+4x_2y_1) \\ & +\lambda (4x_2+4/3.y_1-4/3.x_2y_1) - c \\ & +((w-\lambda^4 (3y_2-3x_1y_2-8/3.y_1+8/3.x_2y_1) - \lambda^3 (-6y_2+6x_1y_2 \\ & +8y_1-8x_2y_1) - \lambda^2 (3y_2-3x_1y_2-8y_1+8x_2y_1) \\ & -\lambda (8/3.y_1-8/3.x_2y_1)) / (\lambda^4 (13/12.-3y_2-3x_1+3x_1y_2 \\ & +8/3.y_1+8/3.x_2-8/3.x_2y_1) + \lambda^3 (-1+6y_2+6x_1-6x_1y_2-8y_1 \\ & -8x_2+8x_2y_1) + \lambda^2 (-0.5-3y_2-3x_1+3x_1y_2+8y_1+8x_2 \\ & -8x_2y_1) +\lambda (-1/3.-8/3.y_1-8/3.x_2+8/3.x_2y_1)+.75) \\ & (\lambda^4 (23/12.-3y_2-3x_1+3x_1y_2+4/3.y_1+4/3.x_2-4/3.x_2y_1) \\ & +\lambda^3 (-3.+6y_2+6x_1-6x_1y_2-4y_1-4x_2+4x_2y_1) \\ & +\lambda^2 (0.5-3y_2-3x_1+3x_1y_2+4y_1+4x_2-4x_2y_1) \\ & +\lambda (1/3.-4/3.y_1-4/3.x_2+4/3.x_1y_1)+.25)) \end{aligned}$$

Equation 5

$$\gamma_o = (w- \lambda^4 (3y_2-3x_1y_2-8/3.y_1+8/3.x_2y_1)- \lambda^3 (-6y_2+6x_1y_2 +8y_1-8x_2y_1)- \lambda^2 (3y_2-3x_1y_2-8y_1+8x_2y_1) -\lambda (8/3.y_1-8/3.x_2y_1)) / (\lambda^4 (13/12.-3y_2-3x_1+3x_1y_2 +8/3.y_1+8/3.x_2-8/3.x_2y_1)+ \lambda^3 (-1+6y_2+6x_1-6x_1y_2-8y_1 -8x_2+8x_2y_1)+ \lambda^2 (-0.5-3y_2-3x_1+3x_1y_2+8y_1+8x_2 -8x_2y_1) + \lambda (-1/3.-8/3.y_1-8/3.x_2+8/3.x_2y_1)+.75)$$

Appendix B

Fortran program for the estimation of λs and γ_os

```
10   REAL RNUM(48), DNUM(48), XLAM(12), XGAM(12)
20   DATA RNUM /1.,1.,.99,.99,1.,1.,1.,2.,1.,1.,1.,1.,
     1.,1.,0.,0.,1.,1.,.99,.99,1.,1.,1.,2.,1.,1.,1.,1.,
     1.,1., 0.,0.,0.,0.,.99,.99,0.,0.,1.,2.,0.,0.,1.,1.,
     0.,0., 0.,0./
30   DATA DNUM /2.,4.,1.,1.,2.,4.,2.,3.,2.,4.,4.,3.,2.,
     4.,1.,1.,4.,8.,1.,1.,4.,8.,2.,3., 4.,8.,4.,3.,4.,
     8.,1.,1., 1.,1.,1.,1.,1.,1.,2.,3.,1.,1., 4.,3.,1.,
     1.,1.,1./
40   READ (50,3,END=300) C,W
50   FORMAT (2F9.5)
60   U=1.-C-W
70   CT=3.*C-W
80   DO 270 ITE=1,12
90   S=RNUM(1+4*(ITE-1))/DNUM(1+4*(ITE-1))
100  T=RNUM(2+4*(ITE-1))/DNUM(2+4*(ITE-1))
110  Y=RNUM(3+4*(ITE-1))/DNUM(3+4*(ITE-1))
120  Z=RNUM(4+4*(ITE-1))/DNUM(4+4*(ITE-1))
130  IF (CT.LE.0.) THEN
```

```
      XM=0.
135    ELSEIF(C.EQ.1.) THEN
      XM=1.
      ELSE Xl=0
      XU=1.
140    DO 220 I=1,20
      XM=(XL+XU)/2.
143    YL=XL**4*(-3.+6*S+3*Z-3*S*Z-4*T-4/3.*Y+4/3.*T*Y)
     +  +XL**3*(4.-12*S-6*Z+6*S*Z+12*T+4*Y-4*T*Y)
     +   +XL**2*(6*S+3*Z-3*S*Z-12*T-4*Y+4*T*Y)
     +  +XL*(4*T+4/3.*Y-4/3.*T*Y)-C
     +  +((W-XL**4*(3*Z-3*S*Z-8/3.*Y+8/3.*T*Y)
     +  -XL**3*(-6*Z+6*S*Z+8*Y-8*T*Y)
     +  -XL**2*(3*Z-3*S*Z-8*Y+8*T*Y)
     +  -XL*(8/3.*Y-8/3.*T*Y)) /
     +  (XL**4*(13/12.-3*Z-3*S+3*S*Z +8/3.*Y+8/3.*T-8/3.*T*Y)
     +  +XL**3*(-1+6*Z+6*S-6*S*Z-8*Y -8*T+8*T*Y)
     +  +XL**2*(-0.5-3*Z-3*S+3*S*Z+8*Y+8*T-8*T*Y)
     +  +XL*(-1/3.-8/3.*Y-8/3.*T+8/3.*T*Y)+.75)
     +  *(XL**4*(23/12.-3*Z-3*S+3*S*Z+4/3.*Y+4/3.*T-4/3.*T*Y)
     +  +XL**3*(-3.+6*Z+6*S-6*S*Z-4*Y-4*T+4*T*Y)
```

```
     + +XL**2*(0.5-3*Z-3*S+3*S*Z+4*Y+4*T-4*T*Y)
     + +XL*(1/3.-4/3.*Y-4/3.*T+4/3.*T*Y)+.25))
145   YM=XM**4*(-3.+6*S+3*Z-3*S*Z-4*T-4/3.*Y+4/3.*T*Y)
    + + XM**3*(4.-12*S-6*Z+6*S*Z+12*T+4*Y-4*T*Y)
    + + XM**2*(6*S+3*Z-3*S*Z-12*T-4*Y+4*T*Y)
    + +XM*(4*T+4/3.*Y-4/3.*T*Y) - C
    + ((W-XM**4*(3*Z-3*S*Z-8/3.*Y +8/3.*T*Y)
    + -XM**3*(-6*Z+6*S*Z+8*Y-8*T*Y)
    + -XM**2*(3*Z-3*S*Z-8*Y+8*T*Y)
    + -XM*(8/3.*Y-8/3.*T*Y)) /
    + (XM**4*(13/12.-3*Z-3*S+3*S*Z +8/3.*Y+8/3.*T-8/3.*T*Y)
    + +XM**3*(-1+6*Z+6*S-6*S*Z-8*Y -8*T+8*T*Y)
    + +XM**2*(-0.5-3*Z-3*S+3*S*Z+8*Y+8*T-8*T*Y)
    + +XM*(-1/3.-8/3.*Y-8/3.*T+8/3.*T*Y)+.75)
    + * (XM**4*(23/12.-3*Z-3*S+3*S*Z+4/3.*Y+4/3.*T-4/3.*T*Y)
    + +XM**3*(-3.+6*Z+6*S-6*S*Z-4*Y-4*T+4*T*Y)
    + +XM**2*(0.5-3*Z-3*S+3*S*Z+4*Y+4*T-4*T*Y)
    + +XM*(1/3.-4/3.*Y-4/3.*T+4/3.*T*Y)+.25))
147    YU=XU**4*(-3.+6*S+3*Z-3*S*Z-4*T-4/3.*Y+4/3.*T*Y)
    +  +XU**3*(4.-12*S-6*Z+6*S*Z+12*T+4*Y-4*T*Y)
    +  +XU**2*(6*S+3*Z-3*S*Z-12*T-4*Y+4*T*Y)
```

```
     +  +XU*(4*T+4/3.*Y-4/3.*T*Y) - C
     +  +((W-XU**4*(3*Z-3*S*Z-8/3.*Y+8/3.*T*Y)
     +  -XU**3*(-6*Z+6*S*Z +8*Y-8*T*Y)
     +  -XU**2*(3*Z-3*S*Z-8*Y+8*T*Y)
     +  -XU*(8/3.*Y-8/3.*T*Y)) /
     +  (XU**4*(13/12.-3*Z-3*S+3*S*Z +8/3.*Y+8/3.*T-8/3.*T*Y)
     +  +XU**3*(-1+6*Z+6*S-6*S*Z-8*Y -8*T+8*T*Y)
     +  +XU**2*(-0.5-3*Z-3*S+3*S*Z+8*Y+8*T-8*T*Y)
     +  +XU*(-1/3.-8/3.*Y-8/3.*T+8/3.*T*Y)+.75)
     +  *(XU**4*(23/12.-3*Z-3*S+3*S*Z+4/3.*Y+4/3.*T-4/3.*T*Y)
     +  +XU**3*(-3.+6*Z+6*S-6*S*Z-4*Y-4*T+4*T*Y)
     +  +XU**2*(0.5-3*Z-3*S+3*S*Z+4*Y+4*T-4*T*Y)
     + +XU*(1/3.-4/3.*Y-4/3.*T+4/3.*T*Y)+.25))
  150   IF(YM.EQ.0.)GOTO230
  160   IF(YL.EQ.0.)GOTO230
  170   IF(YU.EQ.0.)GOTO230
  180   CALL SGN (YU,IU)
  190   CALL SGN (YL,IL)
  200   CALL SGN (YM,IM)
  210   IF(IL.EQ.IM) XL=XM
  220   IF(IU.EQ.IM) XU=XM
```

```
230    IF (YL.EQ.0.) THEN
         XM=XL
       ELSEIF (YU.EQ.0.) THEN
         XM=XU
       ENDIF
       ENDIF
240    G=(W-XM**4*(3*Z-3*S*Z-8/3.*Y+8/3.*T*Y)-XM**3*(-6*Z+6*S*Z
     +   +8*Y-8*T*Y)-XM**2*(3*Z-3*S*Z-8*Y+8*T*Y)
     + -XM*(8/3.*Y-8/3.*T*Y)) / (XM**4*(13/12.-3*Z-3*S+3*S*Z
     + +8/3.*Y+8/3.*T-8/3.*T*Y)+XM**3*(-1+6*Z+6*S-6*S*Z-8*Y
     + -8*T+8*T*Y)+XM**2*(-0.5-3*Z-3*S+3*S*Z+8*Y+8*T
     + -8*T*Y)+XM*(-1/3.-8/3.*Y-8/3.*T+8/3.*T*Y)+.75)
250    IF (G.GT.1) G=1.
260    IF (G.LT.0) G=0.0
       XLAM(ITE)=XM
       XGAM(ITE)=G
270    CONTINUE
280    WRITE (10,30) c,w,u,(XLAM(J),XGAM(J),J=1,12)
290    FORMAT(3F9.5,12(2F9.5))
       GOTO40
300    STOP
```

```
      END
310   SUBROUTINE SGN(VAL,INT)
320   IF(VAL.LT.0.) INT=-1
330   IF(VAL.GT.0.) INT=1
340   RETURN
350   END
```

Appendix C

Explanation of the Program

The finite state scoring polynomial ,equation 4, is a continuous real valued function of a real variable λ, $\lambda \epsilon [0,1]$ and there is a unique solution of λ for any values of c and w. A proof of uniqueness of solution can be seen in García-Pérez (1985). In the program λ, γ_0 are coded as XM and G.

The value of C and W for each examinee were entered the program as input at line 40. Constant term (3C-W) was assigned to variable CT at line 70 to treat examinees whose C values was below or at the guessing level separately in the estimation procedure. Twelve iterations of the first do loop from line 80 to line 270 calculated twelve pairs of estimated λs and γs for each examinee. Within the do loop, line 80 to line 120 generated the values for twelve different pairs of x_i, y_j, assigned these different values in the scoring polynomial and made twelve different finite state scoring schemes. In the program, x_1, x_2 were coded as S and T, y_1, y_2 were coded as Y and Z respectively. The value of λ for examinees whose C value was below or at guessing level were assigned $\lambda = 0$ at line 130. Although none of the examinees in this study got perfect score $C = 1$, the cases with $C = 1$ were taken into account in the program and assigned $\lambda = 1$ for these cases at line 135 without further considerations.

The nested do loop from line 140 to line 220 calculated λ's from the finite state scoring polynomial for examinees whose C values were greater than guessing level and less than perfect. The nested do loop implemented a variant of the bisection

method (Cheney, & Kincaid, pp 76-80) to estimate λ for each examinee for each model. The initial interval was defined as [0,1] at line 141. Variables XL, XU, XM were, respectively, lower extremum, upper extremum, and the mid point of the interval in each iteration. The scoring polynomial was entered to the program at line 143, line 145, and line 147. The error tolerance , the error in the final iteration, e=.000001 was obtained by assigning 20 iterations to the nested do loop at line 140. γ_o's associated with the λ's estimated from all of the above three cases were calculated from the equation 5 in the program at line 240. γ_o values above 1 and below 0 were truncated at line 250 and 260, if these occurred.

VITA

THAN THAN ZIN

Date of birth: 5th October, 1956

Place of birth: Yangon (Rangoon)

EDUCATION

Virginia Polytechnic Institute and State University,

Blacksburg, VA

Ph.D., Educational Research & Evaluation

April, 1992

Virginia Polytechnic Institute and State University,

Blacksburg, VA

Master of Arts, Education, May 1989

Institute of Education, Yangon (Rangoon)

Bachelor of Education, Jan 1981

EXPERIENCE

Office of Measurement and Research Services

Virginia Polytechnic Institute and State University,

Blacksburg, VA

Graduate Research Assistant

1991-present

Computer and Research Laboratory

Administrative and Educational Services, College of Education

Virginia Polytechnic Institute and State University,

Blacksburg, VA

Graduate Research Assistant 1989-1991

State High Schools, Yangon (Rangoon), Myanmar (Burma)

High School Mathematics Teacher

1981-1987

AFFILIATIONS

American Educational Research Association

Mid South Educational Research Association

HONORS Fulbright Scholar (1987-1989)