This is to certify that the

thesis entitled

The Utilization of Antecedent Data in
Conjunction with Test Results for
Curricular Decision Making

presented by

Bernhard Darwin Kaufman, Jr.

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Measurement and
Evaluation

William Mehrens
Major professor

Date February 13, 1980

THE UTILIZATION OF ANTECEDENT DATA IN CONJUNCTION WITH TEST RESULTS FOR CURRICULAR DECISION MAKING

By

Bernhard Darwin Kaufman, Jr.

A DISSERTATION

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

DOCTOR OF PHILOSOPHY

Department of Counseling, Personnel Services
and Educational Psychology

1980

ABSTRACT

THE UTILIZATION OF ANTECEDENT DATA IN CONJUNCTION WITH TEST RESULTS FOR CURRICULAR DECISION MAKING

By

Bernhard Darwin Kaufman, Jr.

Decisions about mastery of an achievement domain are frequently made on the basis of a small sample of items. Because of the small number of items the possibility of incorrect decisions is high. One way of improving these decisions is to utilize additional information in consort with the test information.

This study sought to determine the efficacy of incorporating non-test information into test based decision models. These models were compared, based on classification accuracy. The non-test information variables of the study were instructional time history, instructional testing history, mathematics achievement and sex. The history variables were captured from files maintained on students in a computer managed instructional program. The standard by which the models were compared was mastery classification based on a 156 item test concerning a unit on multiplication and division. This variable also served as the dependent

variable in model development.

There were three phases of analysis in the research. The first used stepwise regression to discover the relationships which existed among the non-test information variables, a set of subtests drawn from the 156 item test and the results of the 156 item test itself. Also, during this phase, the incremental validity of subtests was determined as well as the functional length of subtests combined with instructional time and mathematics achievement.

During phase II least squares and Bayesian models were developed for the purpose of making decisions about mastery of the domain. The least squares model contained mathematics achievement and instructional time as non-test information. In order to apply the Bayesian model, a parameter indicating the value of prior information needed to be set. The coefficient which resulted in the best decision precision established the value of prior information at 2.75 test items.

The final phase compared the Bayesian and least squares decision approaches with the raw score or proportion correct approach for making mastery classifications. Mastery levels of .70, .75, .80, .85, and .90 were examined. None of the approaches stood out as being more effective. Comparison of classification based

the least squares models, containing the non-test information variables with and without a six item subset of the domain, indicated that adding the test information did not improve classification accuracy.

Four conclusions were reached as a result of the analysis. First, a six item test does not improve mastery classification beyond what was possible with pre-existing information. Second, learning rate represents information which is independent of mathematics achievement. Third, neither least squares or Bayesian approaches improve decision precision over that obtained using raw scores. Finally, decision precision is improved when twelve items are used rather than six.

It was recommended that teachers develop ways of using pre-existing information as they monitor pupils. Having measures of achievement and learning rate, one may need only to keep track of on task behavior. Pupils behaviors suggesting frustration can be taken to indicate a need for diagnosis. At such a point, a test of sufficient length to yield accurate decisions can be administered. In sum, if pupils are initially well placed in the curriculum and instructional methods and materials are carefully selected, testing can be restricted to points where diagnosis is indicated by off task behavior reflecting frustration whose cause the teacher cannot

easily identify.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO SYMBOLS AND NOMENCLATURE

π_i: An individual's domain proportion correct

π_o: Mastery level proportion

$\hat{\pi}_i$: An estimate of an individual's domain proportion correct

ω_i: An individual's mastery/non-mastery classification

T : An individual's true score

$\hat{T}$: An estimate of an individual's true score

$\rho_{xx'}$: The reliability of a test

X_i: An individual's raw score

μ_X: Mean raw score for a group

σ_T^2: Classical true variance

σ_E^2: Classical error variance

μ_T: Mean classical true score for a group

σ_X^2: Classical observed variance

γ_i: Arcsin transformation of π_i

g_i: Tukey-Freeman arcsin transformation of X_i

n : Number of items in a test

$g.$: Mean of a group of g_i's

$\hat{\phi}_c$: Classical estimate of true variance where scores have been subjected to Tukey-Freeman transformation

$\hat{\phi}_{gc}$: Classical estimates of observed variance

$\hat{\phi}_{Ec}$: Classical estimate of error variance where scores have been subjected to Tukey-Freeman transformation

$\hat{\gamma}_{ic}$: Classical estimate of arcsin transformation of π_i

χ^{-2}: Inverse chi square

λ: Scale parameter of inverse chi square distribution

ν: Degrees of freedom for inverse chi square distribution

$\hat{\phi}_{b_m}$: Mean of the inverse chi square distribution

t: Test information parameter

$\hat{\gamma}_{ib}$: Bayesian estimate of arcsin transformation of π_i

$\gamma_{\cdot}$: Mean of arcsin transformed π_i's

N: Number of individuals

$\hat{\gamma}_{\cdot b}$: Bayesian estimate of the mean of arcsin transformed π_i's

$\hat{\phi}_b$: Bayesian estimate of true variance where scores have been subjected to Tukey-Freeman transformation

$\hat{\phi}_{Eb}$: Bayesian estimate of error variance where scores have been subjected to Tukey-Freeman transformation

$\hat{\phi}_{gb}$: Bayesian estimate of observed variance where scores have been subjected to Tukey-Freeman transformation

$\hat{\gamma}_{ib_m}$: Bayesian marginal mean estimate of arcsin transformation of π_i

ρ^*: Bayesian marginal mean estimate of proportion of true to observed variance

γ_o: Arcsin transformation of π_o

a_1: Mastery

a_2: Non-mastery

1_{01}: Loss associated with false positive

1_{10}: Loss associated with false negative

EL: Expected loss

μ_i: Mean of posterior marginal

σ_i^2: Variance of posterior marginal

π_1, π_2: Proportion boundaries of the indifference region

X_1, X_2: Raw score boundaries of the indifference region

TEST: Instructional testing history

TIME: Instructional time history

STEP: Sequential test of Education Progress Mathematics Concepts Test

SUBTEST(J): Domain item samples of length J
J = 6, 12, . . .60

DOMAIN: Score on 156 item division and multiplication test

SIX: Classification based on SUBTEST(6)

TWELVE: Classification based on SUBTEST(12)

YHAT: Classification based on least square estimate containing TIME and STEP

YHATP: Classification based on least square estimate containing TIME, STEP and SUBTEST(6)

BAYES6: Classification based on Bayesian marginal mean estimate containing SUBTEST(6)

BAYES12: Classification based on Bayesian marginal mean estimate containing SUBTEST(12)

$\bar{X}_C$: Mean of classifications based on C

$\hat{\Psi}$: Scheffe' contrast

$\hat{\phi}_{\hat{\Psi}}^2$: Variance of a Scheffe' contrast

CHAPTER I

THE PROBLEM

Problem

Individualized instruction requires frequent decisions about each person passing through the curriculum. The basis for these decisions is often an estimate of a domain score based on a small sample of items from the domain. Because of the small sample the possibility of incorrect decisions is great. Millman (1973) has shown that with a mastery level of eighty percent, more than a third of those students whose actual domain achievement is sixty percent will get four of five items correct and thus be misclassified as having mastered the objective or unit.

The test data available in such decision situations is not the only existant information which is pertinent to the decisions. In fact, there is usually information present prior to testing. Cronbach and Gleser (1965) have challenged testers to show that the application of their instruments result in an improvement in the quality of decisions. To use Sechrist's (1963) term, testers should demonstrate the incremental validity of the tests they employ. No such investigation has

been done with domain referenced tests. Thus, it is not known that the estimates of domain scores based on small item samples yield new information for decision making.

Solution

One way of improving the quality of decisions made with the aid of domain test estimates is to utilize additional information in conjunction with the estimate. Such information, once identified may be joined with test information in a mathematical model which should yield improved domain estimates.

There are two statistical approaches to modeling; Bayesian and least square regression. Both of these will likely yield an improved estimate. No research has been done in an applied setting with the domain score known. Therefore there is no empirical basis for recommending one procedure over the other.

Need for the study

Domain tests are being widely used for decision making. It is conceivable that decisions based on short tests alone may be worse than those made knowing only historical information. While it may not be feasible to eliminate tests from an instructional sequence, educators should be alerted to the fact that their

results alone are not a sound basis for decisions. If test data do not provide information, decision makers should be so aware.

Further, if the solutions proposed are sound, this should demonstrate in an applied setting. Then guidance in the application of the procedures should be made available to practitioners.

Purpose of the study

There are two components to the research reported herein. One has to do with the investigation of the information value of several variables, including test data, with respect to results on a domain test. Once these various information relationships were illuminated, two models were compared to each other and a raw score for their efficacy as a bases for criterion referenced decisions. Objectives 1 and 2 below form the first component. Objective 3 the second.

Specifically stated the objectives were:

1. To determine the information existant in four antecedent and collateral variables relative to domain achievement.

2. To couple information with test results in order to determine:
   a. the incremental validity of short domain tests,
   b. if decision precision can be improved by using antecedent and collateral data with test results,
   c. the functional lengths of several short domain tests.

3. to compare the Bayesian marginal mean model, the least square regression model, and the raw score approach with respect to decision precision.

Definition of terms

Given below are definitions of several terms which are used throughout this thesis.

Domain test

"Any test consisting of a random or stratified random sample of items selected from a well defined set or class of tasks."(Millman, 1974, p. 315)

Criterion referenced testing

The use of a test to make decisions about a criterion.

Information

Datum is information if and only if it reduces the uncertainty involved in making a decision.

Functional test length

The length of test necessary to provide information equivalent to that provided by collateral, antecedent and test information.

Incremental validity

The extent to which a multiple correlation is raised by the addition of test results to

a set of prior existing information.

Domain achievement

The proportion of items correct on a set of items which comprehensively cover an objective or set of objectives.

Decision precision

The proportion of correct classifications made on the basis of a given decision algorithm.

CHAPTER II

REVIEW OF THE LITERATURE

Two excellent reviews have been prepared which cover criterion referenced testing comprehensively. These are Millman (1974) and Hambleton, Swaminathan, Algina, and Coulson (1978). Because of the comprehensiveness of these monographs the present review draws heavily on these two papers.

The topics to be covered in this review are: 1) definitions, 2) estimation of domain scores, 3) criterion referenced decisions, 4) validity, and 5) test length. Some of these topics are covered in greater depth than others. The criteria for depth of coverage was the topic's direct relevance to the research. For example, the estimation of domain scores is the direct focus of the study and thus the greatest amount of space is devoted to this area.

Definitions

As Hambleton et al. (1978) have observed there is by no means a single accepted definition of a criterion referenced test. Two quotations which are at opposite poles of the generality continuum illustrate this.

The first is the most restrictive.

> "A pure criterion referenced test is one consisting of a sample of production tasks drawn from a well defined population of performances, a sample that may be used to estimate the proportion of performances in that population at which the student can succeed." (Harris and Stewart, 1971), p. 1)

Ivens defined a criterion referenced test, in most general terms, as one "comprised of items keyed to behavioral objectives." (Ivens, 1972, p. 2) Clearly one must have a referent which is more specifically defined than is the case if both of these quotations are allowed within the class of the concept "criterion referenced test."

The purpose of this section of the review will be to arrive at a term for and definition of the kind of test we are investigating in this research. To do this we will allude to some terms and corresponding referents which will help delimit our concept.

Hambleton et al. (1978) point out that criterion refers to a minimal acceptable level of functioning. This definition is consistent with Glaser and Nitko (1971), Millman (1974), and Harris, et al. (1974). So a criterion referenced test could be one which was used to make a decision about this minimal acceptable level of functioning. Herein lies the problem, when

one applies the accepted definition of criterion; criterion referenced implies only that the test has some relationship to a decision about level of functioning. Looking at it from this point of view, Iven's definition seems most appropriate. That is, a test comprised of items keyed to behavioral objectives defined as Mager (1962) does would be criterion referenced in the sense that the results could be used to make a decision about the minimal acceptable level of functioning.

Glaser and Nitko (1971), consistent with Harris and Stewart (1971), speak of production standard in their definition of criterion referenced but also, as do Harris and Stewart, they use the words "well defined population of performances." So, not only should these tests measure a level of functioning, that level should be generalizable to some larger domain or population. What Harris and Stewart do not allude to is criterion in the sense of minimal acceptable level of functioning.

Hively, et al. (1968), Bormuth (1970) and Osburn (1968) have specified algorithmic procedures for defining a domain of test items. Popham (1975) describes what he calls an amplified objective which specifies in detail the testing situation, response alternatives

and a criterion of correctness, in effect, defining the domain of items. Baker (1974) also provides procedures for carefully defining the item domain of an objective. The direction of the work in this area seems to underline the importance of the notion of domain.

As one might suspect the importance of the domain has motivated the term Domain Referenced Test. Millman (1974) defines such tests as:

> "any test consisting of a random or stratified random sample of items selected from a well defined set or class of tasks." (Millman, 1974, p. 315)

It should be noted that such a definition does not refer to a criterion. The definition of a test can be separated from the specification of a desired level of functioning (as Harris and Stewart's (1971) definition also illustrates). In fact, a single domain referenced test can be used to make decisions about more than one criterion. Admittedly, there is a connection between the decision criterion to be addressed with the results of a domain-referenced test and the definition of the "set or class of tasks." However, in developing the test items the emphasis is on content domain, the criterion can be established separately.

Thus, it seems most appropriate to refer to domain tests. In current practice such tests are most often

used to make decisions about a person's status relative to a criterion. It is appropriate to say that scores are domain referenced and decisions based on the scores are criterion referenced.

The use of the term criterion-referenced testing to describe general approaches whose overall aim is to make decisions about a criterion is useful. Domain or objective referenced tests are but tools which can be employed in this pursuit.

Estimating Domain Scores

The basic problem is; given an individual's observed score on a criterion referenced test, what is his score on the domain, and further, does this represent mastery or non-mastery status (Hambleton and Novick, 1973). To use the symbols which seem to appear most consistently in the literature (Swaminathan, Hambleton and Algina, 1975; Hambleton and Novick, 1973; Novick, Lewis and Jackson, 1973); if X_i (an individual's score is known, what is π_i (the domain score) and further what is ω_i (ω_i=1 if mastery, ω_i=0 if non-mastery). So the problem is to obtain $\hat{\pi}_i$ (an estimate of π_i) and ω_i (an estimate of ω_i).

There are five distinguishable procedures which have been described in the literature for solving this

problem. These are: 1) proportion correct, 2) classical model II, 3) Bayesian model II, 4) Bayesian marginal mean, and 5) the Binomial (Note 1). The first four of these differ from the fifth in that they provide a single direct $\hat{\pi}_i$. The binomial procedure yields information about the probability that π_i is greater than some given mastery level π_o.

The remainder of this section will provide discussion of each of these five procedures.

Proportion Correct

The estimate of the proportion correct is the ratio of correct items to the length of the test. This value can also be thought of as the raw score multiplied by a constant which is the inverse of the number of items. For a small number of items this estimate yields tenuous results. Millman (1974) has shown that for a mastery level of 80 percent, more than a third of those who could achieve only 60 percent of the domain of items will get at least four of five items correct and thus the decision of mastery will be in error. Hambleton, et al. (Note 1) observed that "procedures which take other information into account are more desirable."

Classical Model II

The Classical Model II and Bayesian Model II allow for the inclusion of other information into the decision making process. The classical model includes the mean of the group in which the individual is a member. This is collateral information. The Bayesian Model II considers in addition to the group mean, an investigator's subjective feeling regarding the prior status of the group. The remainder of this section discusses the classical model II in detail.

Jackson (1972) observed that Truman Kelley's (1927) estimate of true score effectively joined test results with the collateral data of group mean. Lord and Novick (1968) state Kelley's formula for the estimate of true score (T) as

$$\hat{T} = \rho_{xx'} X + (1-\rho_{xx'}) \mu_x \qquad (1)$$

Where $\rho_{xx'}$ is the reliability, X test score and μ_x the mean for the group. Thus test data is incorporated through X and the collateral data by way of μ_x. Novick and Jackson (1974) observe that

$$\hat{T} = \frac{\sigma_T^2 X + \sigma_E^2 \mu_T}{\sigma_T^2 + \sigma_E^2} \qquad (2)$$

Classical true score theory (Lord and Novick, 1968) assumes that $\mu_T = \mu_x$. Thus expression (2) can be

rewritten in the form

$$\hat{T} = \frac{\sigma_T^2 X + \sigma_E^2 \mu_X}{\sigma_T^2 + \sigma_E^2} \quad (3)$$

Further, true score theory assumes $\sigma_X^2 = \sigma_T^2 + \sigma_E^2$

so that

$$\hat{T} = \frac{\sigma_T^2}{\sigma_X^2} X + \frac{\sigma_E^2}{\sigma_X^2} \mu_X \quad (4)$$

This expression makes clear the fact that Kelley estimates are "...a weighted sum of two separate estimates, one based upon the individual's observed score X and the other based on the mean of the group to which he belongs..." (Lord and Novick, 1968, p. 65). It can further be observed that when the test is highly reliable (i.e., σ_E^2 is small) the test data is weighted heavily. If the test is not highly reliable then the estimate is more dependent on collateral data namely μ_X.

In order to utilize Kelley's procedure in situations where binary decisions, such as mastery/non-mastery, are to be made; Jackson modified the above procedures. He applied the Tukey-Freeman arcsine transformation to individual scores (X_i) and obtained the transformed estimate

$$g_i = 1/2 \left[\mathrm{Sin}^{-1}(X_i/n+1)^{1/2} + \mathrm{Sin}^{-1}(X_i+1/n+1)\right]^{1/2} \quad (5)$$

Under this transformation of X_i, the corresponding transformed variable γ_i for the proportion correct (π_i) is given by the expression

$$\gamma_i = \text{Sin}^{-1} \sqrt{\pi_i} \qquad (6)$$

If the number of test items is at least eight then the distribution for g_i will be approximately normal with the mean being the transformed value of the proportion correct (γ_i) and the variance $(4n+2)^{-1}$ (Anscombe, 1948). That is $g_i \sim N(\gamma, (4_n+2)^{-1})$. In classical notation this can be written as $g_i \sim N(T, \sigma_E^2)$.

The statement $g_i \sim N(\gamma_i, (4n+2)^{-1})$ is about a fixed person (i) under the hypothetical condition of a finite number of repeated testings. If there is a single testing of a finite number (N) of persons (i.e., i = 1,2...N) then Jackson (1972) has shown that the mean is given by

$$g_\cdot = \sum_{i=1}^{N} g_i/N \qquad (7)$$

and the variance (ϕ_c) is

$$\hat{\phi}_c = [\sum_{i=1}^{N} (g_i - g_\cdot)^2 - (4n+2)^{-1}](N-1)^{-1} \qquad (8)$$

This expression can be rewritten as

$$\hat{\phi}_c = \frac{\sum_{i=1}^{N} (g_i - g.)^2}{N-1} - \frac{N}{N-1}(4n+2)^{-1} \qquad (9)$$

to facilitate the determination of its connection with true score theory. The first term of the expression is the observed variance, the second term is error variance. We can write

$$\hat{\phi}_c = \phi_{gc} - \phi_{EC} \qquad (10)$$

and note that ϕ_c is the analogue of σ_T^2. Also ϕ_{gc} is analogous to σ_X^2 and ϕ_{EC} to σ_E^2.

Returning to (2), the Kelley formula for the transformed variables becomes

$$\hat{\gamma}_{ic} = \frac{\hat{\phi}_c g_i + \phi_{EC} g.}{\hat{\phi}_c + \phi_{EC}} \qquad (11)$$

or

$$\hat{\gamma}_{ic} = \frac{\hat{\phi}_c}{\phi_{gc}} g_i + \frac{\phi_{EC}}{\phi_{gc}} g. \qquad (12)$$

This is clearly a weighted sum of the transformed test scores and the mean of the scores, the mean's weight being inversely related to the reliability of the test.

Once the transformed true proportion correct ($\hat{\gamma}_{ic}$) is obtained, one can return to the original scale by a sine transformation of $\hat{\gamma}_{ic}$, namely

$$\hat{\pi}_i = (1+.5/n)\ \text{Sin}^2\ \hat{\gamma}_{ic} - .25/n \qquad (13)$$

This value ($\hat{\pi}_i$) is the estimated domain proportion correct and is based not only on the proportion correct of a subset of items from the domain but also on a group's performance on the same subset of items.

Bayesian Model II

This model estimate uses test data (X_i), collateral data ($\bar{X}$), as well as prior information. This method requires setting a prior distribution representing an investigator's belief prior to testing and then making revised estimates after testing. These revised estimates are based on prior beliefs as well as an individual's test results and the group mean. The distribution which takes all three pieces of information into account is called the posterior distribution.

The question of determining the correct prior distribution has been the subject of considerable theoretical study by Novick and his colleagues (Novick et al., 1973 and Swaminathan, et al., 1975). The current status of these investigations suggest the following.

a) The specification of the mean is not particularly important and may be represented by a uniform distribution in which any score is equally likely.

b) The prior beliefs about variance can be adequately represented by an inverse chi square distribution with two parameters; scale and degree of freedom.

i) The degree of freedom parameter (ν) should be set at 8.

ii) The scale parameter (λ) can then be solved for in an equation with a single unknown namely, the variance. The equation is

$$\lambda = (\nu-2)\hat{\phi}_{b_m} \qquad (14)$$

iii) The necessary estimate of the variance ($\hat{\phi}_{b_m}$) can be obtained as follows:

a) Specify the true proportion correct for the typical examinee in the sample.

b) "...specify the number of test items, t, that would have to be administered to the examinee in order to obtain as much information about π_i as is deemed to be available (Note 1, p. 31).

c) $\hat{\phi}_{b_m}$ is then defined by the equation

$$\hat{\phi}_{b_m} = (4t+2)^{-1} \qquad (15)$$

d) The true proportion correct (γ_{ib}) is then estimated by

$$\hat{\gamma}_{ib} = \frac{g_i\left[\frac{\lambda+\Sigma(\gamma_i-\gamma_\cdot)^2}{N-\nu-1}\right] + \gamma_{\cdot b}(4t+2)^{-1}}{\frac{\lambda+\Sigma(\gamma_i-\gamma_\cdot)^2}{N-\nu-1}] + [4t+2]^{-1}} \quad (16)$$

and the mean of the proportion correct $(\gamma_{\cdot b})$ by

$$\hat{\gamma}_{\cdot b} = \frac{\Sigma\gamma_i}{N} \quad (17)$$

e) Novick et al., (1973) observe that this is equivalent to

$$\hat{\gamma}_{ib} = \frac{g_i\,\hat{\phi}_b + \gamma_{\cdot b}\,(4t+2)^{-1}}{\hat{\phi}_b + (4n+2)^{-1}} \quad (18)$$

where $\hat{\phi}_b = (N+\nu-1)^{-1}[\lambda+\Sigma(\gamma_i-\gamma_\cdot)^2]$ (19)

$\hat{\phi}_b$ is the Bayesian true variance estimate for $\hat{\gamma}_{ib}$, ϕ_{Eb} is the Bayesian error variance estimate for $\hat{\gamma}_{ib}$, and $\hat{\phi}_{gb}$ is the Bayesian observed variance estimate for $\hat{\gamma}_{ib}$. Using this notation (16) can be rewritten as

$$\hat{\gamma}_{ib} = \frac{\hat{\phi}_b}{\phi_{gb}}\, g_i + \frac{\phi_{Eb}}{\phi_{gb}}\, \gamma_{\cdot b} \quad (20)$$

As Novick et al., (1973) indicate, this estimate has a form analogous to Kelley's true score estimation

procedure. The differences between $\hat{\gamma}_{ib}$ as estimated by equation (18) and $\hat{\gamma}_i$ as estimated by equation (12) result from the procedures used for determining the several variance components and the use of $\gamma_{\cdot b}$ as the true mean rather than g.. Table 1 allows a comparison of the Bayesian and Classical variance estimates. Examination of the formuli in the table indicates that prior information is incorporated into the estimate of $\hat{\gamma}_b$ by the estimation procedure for $\hat{\phi}_b$. λ is determined by

$$\lambda = (\nu - 2)(4t + 2)^{-1} \tag{21}$$

where t is the number of test items that would need to be administered to the examinee to obtain as much information about π_i as is deemed available prior to testing. Further, because of the iterative nature of the solution of equation (16), the $\gamma_{\cdot b}$ obtained for the concluding iteration will have been influenced by the value of t.

Thus differences in estimated values for γ are a function of differing amounts of regression due to the variance estimates as well as a different "true" mean on which the regressions occur. Theoretically, the advantage of the Bayesian Model II procedure rests on an improvement in the estimates of true variance,

Table 1. A comparison of Bayesian and Classical variance estimates

Model	BAYESIAN	CLASSICAL
Variance		
TRUE	$\hat{\phi}_b=(N+\nu-1)^{-1}[\lambda+\sum_{i=1}^{N}(\gamma_i-\gamma.)^2]$	$\hat{\phi}_c=\frac{\Sigma(g_i-g.)^2}{N-1}-\frac{N}{N-1}(4n+2)^{-1}$
ERROR	$\phi_{Eb}=(4t+2)^{-1}$	$\phi_{EC}=\frac{N}{N-1}(4n+2)^{-1}$
OBSERVED	$\phi_{gb}=\hat{\phi}_b+\phi_{Eb}$	$\phi_{gc}=\frac{\sum_{i=1}^{N}(g_i-g.)^2}{N-1}$

observed variance and the true mean accomplished by incorporating prior information through the parameter λ.

Bayesian Marginal Mean

Lewis, Wang and Novick (1973) observe that if one wishes to make overall decisions about all groups, joint estimates such as those of Bayesian Model II are appropriate. However, they note that for individualized instruction, decisions about each individual are usually desired and therefore marginal estimates are indicated.

Hambleton, et al., (Note 1) note that the Bayesian Model II requires complicated iterative solutions. Tables prepared by Wang (1973) allow relatively easy computation of marginal estimates. The procedure demands that the degree of freedom parameter be set (again 8, according to Novick, et al. (1973)) and ϕ_{ib_m} is determined by specifying t in the manner described above. With these values ρ^* can be read from Wang's table and the estimate of $\hat{\gamma}_{ib_m}$ is

$$\hat{\gamma}_{ib_m} = g. + \rho^* (g_i - g.) \qquad (22)$$

which can then be transformed to π_i by equation (13).

The marginal mean procedure is an extension of the Bayesian Model II and as such effectively considers the three types of data; test, collateral, and prior

beliefs. It should be understood that all of the Bayesian estimates have been designed for use when one's knowledge of prior status can at best be represented by subjective belief about t. It is this subjective belief which is quantified by the method described for establishing the Bayesian true variance.

The parameter ρ^* is an estimate of

$$\frac{\hat{\phi}}{\phi + \phi_E}$$

a reliability indicator. Lewis, Wang and Novick (1973) report that an empirical study of

$$\rho = \frac{\hat{\phi}_b}{\hat{\phi}_b + \hat{\phi}_{Eb}}$$

and ρ^* indicate that "...ρ^* is substantially larger than ρ for moderate n." (p. 12) As the number of items increase, the discrepancy between ρ and ρ^* becomes smaller (p. 13), and thus estimates of $\hat{\gamma}_b$ and $\hat{\gamma}_{b_m}$ become increasingly similar.

One might expect that if the Bayesian methods do allow for a meaningful incorporation of prior information into the computation of ρ, then these values would be larger than for the corresponding classically computed values. However, in at least one empirical study this was not the case (see Novick et al., 1973, pp. 39-41)). In this instance, the investigators

questioned the estimates of $\hat{\phi}_i$. It would seem that dilemmas such as this are best addressed by studying the quality of decisions made by various estimates.

Binomial Model

One may use the binomial model discussed by Millman (1974) for making probability statements about the true achievement status of an individual. In order to do this three parameters are needed; minimum passing score, number of items and the level of certainty required for establishing mastery.

With these values specified, mastery/non-mastery decisions can be made to a prescribed probability level knowing only the actual score on a test. Tables prepared by Millman (1972) make this model very simple to apply.

As Millman (1974) has observed, all Bayesian approaches yield a regressed estimate of domain scores. That is, if an individual's obtained score is below the group's mean, her estimated domain score will be higher than her obtained score. Analogously, if her obtained score is above the mean, her estimated score will be lower. These statements also hold for classical model II. Such statements do not hold for Millman's binomial model.

Criterion Referenced Decisions

Only Hambleton, Novick and their colleagues (Hambleton, et al., 1973; Hambleton, 1974; Swaminathan et al., 1975; Hambleton, et al., 1978) seem to have given attention to the problem of making decisions based on domain estimates. It will be recalled from the previous review of procedures for estimating domain scores that once π_i is obtained one must determine the appropriate value of ω_i. In the binary classification; if $\gamma_i \geq \gamma_o$ then $\omega_i = 1$ or if $\gamma_i < \gamma_o$ then $\omega_i = 0$. Both γ_i and ω_i are true values and in practice must be estimated. Hambleton et al., (1978) and Swaminathan et al., (1975) have presented a method for ascertaining $P(\omega_i = 1)$ on the basis of Bayesian posterior distributions.

Whenever a decision is made in the face of uncertainty, there will be misclassification. In the case of mastery/non-mastery classification there are two decision actions available. We will call them a_1 (mastery) and a_2 (non-mastery). In this binary case there are also two kinds of error to be made. If the action is a_1 and $\omega_i = 0$, this is called a false positive error. If the action is a_2 and $\omega_i = 1$, this is referred to as a false negative. With each error type some loss is incurred. In the testing setting these may

be unnecessary consumption of materials, teacher time, or student affect such as boredom or frustration. The loss associated with false positives can be symbolized by $L(\omega_i = 0, a_1) = 1_{01}$. The false negative loss is given by $L(\omega_i = 1, a_2) = 1_{10}$. The aim of decision classification is to minimize the expected loss (EL) associated with the action. Thus $EL(\omega, a)$ is to be minimized. The two loss functions are:

$$EL(\omega, a_1) = 1_{10} P(\gamma_i < \gamma_o) \text{ and} \tag{23}$$

$$EL(\omega, a_2) = 1_{01} P(\gamma_i \geq \gamma_o). \tag{24}$$

The decision rules are:

$$a_1 \text{ if } 1_{01} P(\gamma_i \geq \gamma_o) < 1_{10} P(\gamma_i < \gamma_o) \tag{25}$$

$$a_2 \text{ if } 1_{01} P(\gamma_i \geq \gamma_o) > 1_{10} P(\gamma_i < \gamma_o) \tag{26}$$

$$\text{if } 1_{01} P(\gamma_i \geq \gamma_o) = 1_{10} P(\gamma_i < \gamma_o) \tag{27}$$

one is equally well off with either decision.

No one appears to have tackled the problem of estimating 1 for the two forms of misclassification possible in mastery/non-mastery decisions. Hambleton (1974) has speculated on the matter. He feels that false positive error is more serious than false negative error since a student will have a second chance in most systems. Further, if the subject matter is

hierarchical a false positive will likely be frustrated by attempts to achieve future objectives.

Hambleton et al., (1978) describe procedures for determining the probability of mastery given a domain score for each of the Bayesian models reviewed earlier. First one must calculate the mean (μ_i) and variance (σ_i^2) of the posterior marginal distribution (i.e., posterior for each score) by using formulas given in Note 1 . Then a z score is calculated for each individual by $z = \gamma_o - \mu_i/\sigma_i$. This result can then be used with any table of normal deviates to find the probability that an individual's π_i is above the matery level (π_o).

The final step is to combine loss values with probabilities of mastery ($P\ (\gamma_i \geq \gamma_o)$) and non-mastery ($P\ (\gamma_i < \gamma_o)$). By comparing

$$EL\ (\omega, a_1) = 1_{01}\ P\ (\gamma_i \geq \gamma_o) \text{ with}$$

$$EL\ (\omega, a_2) = 1_{10}\ P\ (\gamma_i < \gamma_o).$$

and taking the action corresponding to the smaller of the two one makes the decision with the smallest expected loss.

All of this work is theoretical. No reports of attempts to find actual values for 1 have been published. Such investigations are necessary. In the

absence of supported rationale to the contrary, the practice of setting $1_{01} = 1_{10} = 1$ seems most sensible.

Validity

Cronbach (1971) states:

> "The phrase validation of a test is a source of much misunderstanding. One validates, not a test, but an interpretation of data arising from a specified procedure." (p. 447)

The same authority also points out that there are "... two uses of tests; (a) for making decisions about people tested and (b) for describing these people (p. 445). In the criterion-referenced testing situation where domain tests are employed, it seems that these two uses suggest three validity questions for domain tests utilized for criterion referenced decisions. (1) Is the test content valid? (2) Is the test domain valid? (3) Is the test criterion valid? Hambleton et al., (Note 1) argue that the question of content validity is inextricably twined with domain specification and thus the validity of the content is a function of the adequacy of these specifications. Two procedures for systematically specifying the content domain are item form and amplified objectives (Millman, 1974).

> "An item form has the following characteristics: 1) it generates items with a fixed syntactical structure; 2) it contains

> one or more variable elements; and 3) it defines a class of item sentences by specifying the replacement sets for the variable elements" (Osborn, 1968, p. 97)

Such procedures seem best suited for mathematical and scientific content areas. The second procedure which Hambleton et al., (Note 1) believe:

> "provides an excellent balance between the clarity achieved with item generation schemes and the practicality of behavioral objectives" (p. 15)

are what Popham (1975) calls amplified objectives. These are

> "...expanded statements of an educational goal which provides boundary specifications regarding testing situations, response alternatives, and criteria of correctness." (Millman, 1974, p. 335)

In fact an amplified objective can contain an item form as defined by Osborn. While both approaches are tedious, Popham's approach would seem to provide a means of overcoming Ebel's (1971) concern that only trivial domains can be specified.

The question of domain validity can be answered, as Millman (1974) observes, by determining the relationship between scores on tests X and Y when X is composed of a randomly selected set of items from Y, the set of all items in the domain.

To answer question (3) one must determine the adequacy of criterion related decisions based on the

domain score or an estimate of same. Within the mastery learning model such decisions are about mastery/non-mastery. In such a setting one might think that domain and criterion validity would be the same since the decision directly relates to the content. However, since small differences near the decision threshold are less significant for domain validity than for criterion validity, these values may differ markedly.

One should note that a domain test is itself a procedure involving item specification rules, items, and sampling plans. Domain and content validation determine the adequacy of these procedures. However it seems that these two types of validation are independent of interpretation of results or estimates of the domain. Certainly, criterion validity seeks to address interpretation issues, namely the adequacy of decisions based on a test's results.

Sechrist (1963) used the term Incremental validity to refer to the extent to which a variable raises the multiple correlation when it is included in a set of predictor variables. Cronbach and Gleser (1965) suggested that testers justify the use of instruments by showing that an improvement in some decision resulted from application of tests and further that the magnitude of improvement warranted the cost involved. Thus they

were challenging testers to determine the incremental validity of their instruments. In the context of criterion referenced testing, if prior information is available, does the data in the form of an estimate of the domain score actually reduce uncertainty below the level possible with the prior data alone? The issue of incremental validity is a special case of the criterion validity question in that it relates to decisions based on estimates of the domain score.

Domain Test Length

As Hambleton et al., (Note 1) observe;

> "The problem of determining test length (in the criterion-referenced situation) is related to the size of the misclassification errors one is willing to tolerate." (p. 63)

In general, the longer the test the smaller the size of misclassification error. However, the reality of objective based curriculum systems which use a number of domain tests to make criterion decisions is that the feasible length of tests is quite restricted. Novick and Lewis (1974) feel that twenty items per objective is too large. However, in practice, criterion decisions are often made based on results of tests of five or fewer items.

Two avenues have been explored as means of specifying test length given the specific magnitude of misclassification error one is willing to tolerate. Novick and Lewis (1974) have developed the Bayesian solution to the problem while Millman (1972) and Fhaner (1974) have used strict binomial methods. Each of the two approaches will be discussed in the remainder of this section.

Millman's procedure yields the proportion of misclassifications given an examinee's true proportion correct (π_i), a mastery level (π_o) and the number of items on a test. By applying the tables (Millman, 1972) one can determine the number of items necessary to make a decision which is accurate to a given probability level. A disadvantage is that to use the method one must have the true score if specific recommendations about test length are to be made. Millman's tables are important in that they show, theoretically, the high degree of uncertainty in making criterion decisions based on curriculum embedded tests. For example, the probability of a false negative when $\pi_o = .8$ and $\hat{\pi}_i = .8$ and $n = 15$ is .35.

Fhaner (1974) model is based on the same binomial theory as Millman's. One must specify an indifference region ($\pi_1 < \hat{\pi}_i < \pi_2$) about the cutoff score within which

classification errors are considered unimportant. Then, with the acceptable probabilities of misclassification specified, the necessary test length and X_1 and X_2 corresponding to π_1 and π_2 may be solved for using a normal approximation to the binomial. Unfortunately, using this procedure if $\pi_1 = .70$ and $\pi_2 = .80$, then $n = 121$: clearly an unacceptably long test. Looking at the problem in reverse, Fhaner's empirical investigations show that if between 12 and 17 items are used $\pi^2 = \pi^1 = .3$. This seems quite large for most applications. Simply, both binomial models suggest that tests of the typical length used in criterion referenced situations will lead to many misclassifications. The notion of effective lengthening of tests by utilizing prior knowledge is hopeful.

Novick et al., (1973) suggest that the kind of subjective prior information they foresee being used as worth between six and fifteen additional items. Coupled with their recommendations that a test length of twelve or less is "very desirable" (Novick and Lewis, 1974, p. 158), it is suggested that the application of Bayesian methods outlined in the earlier section on estimating domain scores could result in reasonable length testing sessions and decision certainty equivalent to that achieved with tests containing eighteen to thirty-seven items.

Novick and Lewis (1974) have explored a Bayesian model with the prior distributions having specific means. They observed that when "...the average test score of the group is high (i.e., above the criterion level) and there is little variation among individuals, shorter tests become feasible." (p. 148) These authors have developed tables which yield test length and minimum X_i for advancement. These tabular recommendations are based on prior distributions with known mean. Values are recommended for several loss ratios. In addition one is able to take into consideration his feelings about the extent of dispersion in the prior. Hambleton et al., (Note 1) caution that the recommendations hold for the Bayesian Beta Binomial Model only and the optimality of the recommendations for the Bayesian models reviewed earlier is not known.

Summary

It was concluded from the review of definitional issues about criterion referencing and domain referencing that decisions are criterion referenced and tests are domain or objective referenced.

The literature reflects a concern about using proportion correct or raw score for criterion referenced decisions when only a few items are present. Several

scholars have sought to develop effective procedures which allow information in addition to test scores to be considered in the decision algorithm. There has been investigations of two approaches for accomplishing this.

The first is Truman Kelley's classical true score model. This approach makes use of information about the status of the group from which individuals are drawn to arrive at the best estimate of an individual's standing on the variable being measured. The extent to which group status is considered is a function of the proportion of true variance accounted for by the test. Since domain or objective referenced tests often yield scores based on a few items and thus do not account for the desired level of true variance, many applications of this model to criterion referenced situations will result in the incorporation of collateral data, namely, the group mean, into the estimation process.

A second procedure for adding non-test information to the estimation process follows the Bayesian model. This methodology also takes group mean into the estimation. In addition, this approach attempts to incorporate antecedent information (t) by asking the investigator to set the number of items which would provide information equal in amount to that which he has about the subjects

prior to testing. As with Kelley's approach, the extent of inclusion of the group mean is a function of the amount of error variance. This error factor is a single valued function of t.

Some attention has been given to the theoretical questions of making decisions based on domain estimates. An algebra for incorporating loss values into the decision process has been developed, however, this writer could find no work which provided insights into the problem of how to best set loss values.

Content, domain and criterion validity have been delineated in this chapter. The first two are primarily dependent on the adequacy of content or domain specifications. The domain validity of a subset of items which meets specifications is the correlation of results on that subset with results over the entire domain. Criterion validity reflects the precision with which decisions are made about reaching or not reaching a standard.

Another validity issue addresses the question of whether test information improves decision making and if so, to what extent. This issue is an important one for those contemplating use of short domain or objective referenced tests.

The final section of this chapter addressed research on domain test length. The use of the binomial model to make decisions about test length were discouraging. The binomial literature suggests a need for tests much longer than seem practical in most instructional applications. The use of antecedent and collateral data may help reduce test length requirements.

The research which is described in the following chapters aimed to determine if decision precision can be improved by use of antecedent and/or collateral information. The research proceeded in three stages. In the first, information present in several antecedent and collateral variables was determined. Then, least squares and Bayesian domain estimation models were developed. Finally, the mastery/non-mastery classifications based on the least squares, Bayesian and raw score (proportion correct) approaches were compared.

CHAPTER III

DESIGN AND PROCEDURES

Population

The population for this study was all fifth grade students involved in the MICA (Managed Instruction with Computer Assistance) project in the Madison (Wisconsin) Public Schools during the year of 1976. Six of the thirty-three elementary schools in the district participated in the project. Of the six, four were located on the east side of the cities' isthmus and two on the west side. Madison has a high percentage of professional and white collar workers. The west side of town is primarily residential with some large business enterprises such as insurance companies and financial institutions. The east side of Madison contains the city's industry. The oldest residential areas are on the east side and, in general, property values are lower in the eastern part of the city. Most blue collar workers live on the east side.

The schools which participated in the MICA project were selected primarily because of the interest of their staffs and administrators computer managerial instruction. The final judgment regarding which schools were selected was made by the MICA project director using his

knowledge of the teachers and principals of the schools which were interested in joining the project.

Sample

Four of the six schools in the project were selected for the study. Both schools on the west side were chosen along with two of the four on the east side. These two were picked at random. The four schools had a total of 225 students in the MICA mathematical project in the fifth grade. For the purpose of this study values for all of the variables defined in the next section needed to be available for each subject. This was the use for 172 of the 225 students. These 172 comprised the sample for the study. Since a portion of the research required a cross validation group, the sample was randomly dichotomized into sub samples of seventy-five and ninety-seven. The smaller of the sub samples was used for cross validation purposes.

Variables

Each subject took a one hundred fifty-six item test which covered material on a single instructional unit called Introduction to Multiplication and Division. This unit was comprised of six objectives. The test contained twenty-six items for each of the objectives.

For each student in the two groups the student history file of the MICA system was queried to amass testing history, number of school days of instruction per objective, and sex. The student history file was automatically maintained and updated by MICA software. Each time a student took a test the score was inputed by computer terminal. Records of the date of initial pretesting and successful post testing were kept by the system. Thus accurate information about testing and rate of objective achievement were available for each subject in the MICA project. In addition, standardized achievement results were gathered for each child.

The variables of the study can be divided into three types: those containing information about domain achievement, the actual domain achievement, and those containing information about decisions (mastery/non-mastery) about domain mastery.

The information variables are:

1. Instructional Testing History (TEST)
2. Instructional Time History (TIME)
3. Sequential Test of Educational Progress (STEP)
4. Sex (SEX)
5. Domain Item Samples (SUBTEST (J))

The measure of actual domain achievement was a 156 item test (DOMAIN).

The decision variables were designations of mastery or non-mastery on the basis of several decision criteria.

The criteria were:

1. Score on SUBTEST (J), 12, 18, ... 60
2. A least squares estimate of DOMAIN containing a subset of the information variables but not SUBTEST (J).
3. A least squares estimate of DOMAIN containing SUBTEST (J), J=6 and 12.
4. A Bayesian estimate of DOMAIN, based on SUBTEST (J), J=6 and 12.

A definition of each of the variables is written below.

1. Instructional Testing History (TEST):

This variable was the mean number of instructional tests per objective taken during the four month school period preceeding testing on the large multiplication and division domain test.

2. Instructional Time History (TIME):

This variable was the mean number of school days that each subject in the sample spent per objective during the four month school period preceeding testing on the domain test.

3. Sequential Test of Educational Progress (STEP):

This variable was the raw score on the STEP Mathematical Concepts Test taken prio to domain testing.

4. SEX

This variable was the sex of the subject.

5. Domain Item Samples (SUBTEST (J)):

There were several variables in this category. Each was the raw score on a sample of items

drawn from the items in the domain. Ten such subtests of length 6, 12, ...60 were created. Each was the result of stratified random sampling (without replacement) from the 156 items. The stratification factor was objective. Since the sampling process was done separately for each of the J subtests, it was possible for an item to be present in more than one of the subtests.

6. Domain Achievement (DOMAIN)

This variable was the raw score on the 156 item unit Test. This test was comprised of 156 items covering a unit on multiplication and division. The unit contained six objectives for which there were 26 items each. The items on the domain test were typically used as pre, post and review tests in the MICA system. The development of these items began in the late 1960's. They were written by teachers and underwent continual content review. Since 1972, when the items became part of the MICA math program, the items have undergone analysis to assure they were keyed correctly and that the foils were plausible. In addition the content validity was again assessed by members of the MICA project staff.

Each of the classification variables was calculated for five mastery levels: seventy, seventy-five, eighty, eighty-five and ninety percent. The proportion corresponding to each mastery level was multiplied by the length of the test and this figure was rounded to the nearest integer to obtain the various criteria. Then the values of the classification criteria were compared with individual subject values to obtain the classification variable values. These were zero of one in all cases. There were six bases for classification, which are defined below.

7. Mastery or non-mastery based on the six item domain sample (SIX).
8. Mastery or non-mastery based on the twelve item domain sample (TWELVE).
9. Mastery or non-mastery based on a least squares estimate which contains the variables TEST, TIME and STEP. The development of this variable is discussed later in this chapter (YHAT).
10. Mastery or non-mastery based on a least squares estimate composed of those variables listed for (9) plus SUBTEST (6). (YHATP).
11. Mastery or non-mastery based on a Bayesian estimate which utilizes the information of the variables listed for (9) above and is based on SUBTEST (6) and SUBTEST (12). (BAYES6 and BAYESI2)

Several descriptive statistics were calculated for the variables of the study. Means, standard deviations and ranges were calculated for all variables. In addition, Hoyt reliabilities and standard errors of measurement were obtained for DOMAIN and SUBTEST (J) (J=6, 12,...60) as well as the six objectives of the domain test.

Methodology

There were three stages of analysis for this study. The first was undertaken to discover the information relationship which existed between the antecedent and concommitant variables and the domain achievement. The second was to develop least squares and Bayesian models for making estimates about domain achievement. In the third phase, classifications based on the estimates of the models developed in Phase II were compared. The remainder of this section will discuss the methods used in each of the three phases.

Phase I

The approach used in this phase is based on the assumption that data represents information if and only if it reduces uncertainty involved in making a decision. Least squares stepwise regression (Draper and Smith (1966), Kerlinger and Pedhauzer (1973), Rose-

boom (1966), and Cohen and Cohen (1975)) was employed to reveal the information which existed between STEP, TEST, TIME, SEX and SUBTEST (J) (J=6, 12,...60).

The first step in determining the information value of the information variable was to examine the correlation matrix of these variables and DOMAIN. This was followed by a series of stepwise regression analyses to determine the most parsimonious set of antecedent variables which contained information about

Specifically, the partial coefficients of correlation and alienation were examined. Partial correlation provides an index of the amount of information in one variable (say X) relative to a second (say Y) which is distinct from that information present in yet a third variable (call it W) relative to the second (Y). Conversely, the partial coefficient of alienation is an index of the uncertainty present when a decision about one variable (Y) is made on the basis of the information in a second variable (X) which is distinct from that in yet a third variable (W).

Phase one identified the interrelationships between the variables relative to domain achievement. The zero, first and second order partials for the antecedent and concommitant variables relative to the criterion were examined for the purpose of describing the infor-

mational relationships. Based on this examination and the t-tests of significance for the various partial regression weights those information variables which contain information relative to domain achievement were identified. Then a regression equation was developed. All of the partial regression weights of this equation were significant at the .05 level. The partial correlations indicated the distinct information present in each variable. The coefficient of alienation ($1-R^2$) was the proportion that uncertainty was reduced by considering the set of independent variables as information about domain achievement.

The concepts of incremental validity and functional length of a test were used to illuminate the value of the prior information relative to test information. Incremental validity was determined by adding domain item samples of ascending size (SUBTEST (J) J=1,6,12,... 60) to the regression equation while at the same time deleting the sample of the previous size (SUBTEST (J-1) with J≠1). The incremental validity was the change in R^2 when each domain item sample was considered with the prior information. The t-test of the regression weight for the item sample variable indicated if the incremental validity was significantly different from zero. The difference between the coefficient of alie-

nation prior and subsequent to adding item sample results indicated the degree of uncertainty reduction in the domain estimate accomplished by use of test results.

The incremental validity of three tests of lengths 6, 12, and 18 items were assessed. The tests were generated by stratified random sampling from the original domain of one hundred and fifty-six items. Stratification was based on the six objectives. Sampling was with replacement of items to the domain pool after each test was generated so that items could appear on more than one test. In effect, these equations were studied in a stepwise fashion. The base variables were those found to be significant in the earlier portion of Phase I. The stepped in variable was, in each case, the domain item sample raw score. The regression weight of each test was evaluated using an F-test. Multiple correlations and coefficients of alienation were calculated for the equation containing only the base variables and also for each of the three equations with test results added. This allowed determination of 1) the existence of incremental validity, 2) the magnitude of existent incremental validity, and 3) the extent of uncertainty reduction attributable to the test results.

The functional lengthening of a test refers to a process of utilizing prior data in conjunction with

test results for improving the quality of decisions. When three types of data were incorporated into a decision model, the functional length of the test when coupled with the other data was equal to the length of the hypothetical test which provided the same amount of information when used alone.

The functional length of the 6, 12 and 18 item subtests were investigated. Stratified random item samples of the respective lengths were generated. Then three equations were constructed; each contained the results of one of the three short tests and the base variables of the equation developed earlier. The functional length of a test is defined as the length of the test whose correlation with the domain is equal to the multiple correlation of the model containing the information variables and the results of SUBTEST (J). So if the correlation between SUBTEST (12) and DOMAIN is equal to the multiple correlation of the model containing the information variables and the results of SUBTEST (6), then the functional length of SUBTEST (6) is twelve. Linear interpolation was used to find functional length for the intervals between 6 and 12 and 12 and 18.

Phase II

The purpose of this phase of the analysis was to develop least squares and Bayesian models in order to make decisions about mastery and non-mastery of the domain. The development of the least squares model followed directly from the work done in phase I.

Least squares model development aimed to identify the most parsimonious set of variables for predicting domain achievement, namely the domain score. The desired model has the form of a linear equation whose partial coefficients were each significant at the .05 level. Various permutations of the variables were considered in a stepwise fashion. This assured that the final equation was parsimonious while at the same time containing the maximal information about the criterion variable.

A second least square model was developed by simply adding SUBTEST (6) to the information model.

To assure that the errors of estimation would not be correlated with the independent variable, the least squares statistics were based on a different sample than the one to be used in phase III. The model building sample contained 97 subjects drawn at random from the 172 subjects sample of the project.

The Bayesian model used in this study is the

model II which was derived by Novick, Lewis and Jackson (1973) and Lewis, Wang and Novick (1973) and discussed by Millman (1974) and Hambleton, et al., (1978). The Domain Score estimate was

$$\mu_j = \rho^* g_j + (1-\rho^*) g. \qquad (30)$$

where

$$g_j = \text{Sin}^{-1}[(x_j + 3/8)/(n + 3/4)]^{1/2}, \quad (x_j$$

and n representing j's raw score and the number of items respectively).

$$g. = \sum_{j=1}^{m} g_j/m \quad (m \text{ is the number of subjects})$$

and ρ^* is the Bayesian estimate of the proportion of true to observed variance.

The statistic ρ^* is a function of the prior information about the sample expressed as the length of the test whose sum of correct responses would represent the same amount of information as was available without testing.

In order to get some empirical notion of the effect of this information factor on classification, ρ^* was calculated for 62 values of t ranging from 2.75 to 18 in increments of .25.

Based upon the fluctuation of the classification variable as a function of t, the decision about which

t values would be used in phase II of the analysis was made. Another factor considered the decision process leading to the selection of the precise Bayesian equations to be used for phase III was the data obtained in phase I analysis pertaining to incremental validity. Based on the point of view of Millman (1974) and Swaminathan, et al., (1975) the most appropriate t value is the number of test items which would yield as much information as was available without testing. From this, one would deduce that the value of the classification variable should reach a maximum at about the points where t equals the test length value of the prior information.

Phase III

Phase III of the analysis focused on the comparison of several approaches for making decisions about mastery or non-mastery of this achievement domain. There were three specific approaches studied. These were referred to as: 1) the raw score approach, 2) the least square approach and 3) the Bayesian approach. Within the raw score approach, scores on SUBTEST (6) and SUBTEST (12) were the criterion for classification. Within the least square approach two models were used. The independent variables of the first were STEP and TIME. The second model used these two independent variables

as well as SUBTEST (6). Two Bayesian models were examined: one based on SUBTEST (6) and a second based on SUBTEST (12) Both attempted to incorporate information of the decision variables through specifications of t. Thus, in all, there were six separate models for making classification decisions.

Classifications were made at five mastery levels for each of the models. As a result there were a total of 30 decision criteria, five for each model. The classifications made on the basis of each of the 30 criteria were compared to the mastery - non-mastery classification made on the basis of the 156 item domain score. These comparison resulted in 30 unique vectors of 75 zeros and ones. Each element of a vector represented a subject. The value of the element indicated concordance of model classification and domain classification.

These data were represented as a two-way design with repeated measures on each of the factors. The fixed levels of Factor 1 were the six models. The five mastery levels comprised the levels of Factor 2. The variance was analyzed to determined if there were model effects or mastery level effects. The two F tests for main effects were:

$$F_{mastery} = \frac{MS_{mastery}}{MS_{within}} \quad \text{and}$$

$$F_{model} = \frac{MS_{model}}{MS_{within}}$$

Under the assumption of homogeneous correlations of all pairs of levels on each fixed factor these values have F distributions with 4 and 20, and 5 and 20 degrees of freedom respectively. If the homogeneity assumption is not met, the most conservative F distribution of the ratios will have 1 and 5 and 1 and 4 degrees of freedom respectively. (Greenhouse and Geisser, 1959).

In this study, the homogeneity assumption was not tenable. Thus, the most conservative F distribution was used to determine if there were mastery or model effects at the .05 level. If calculated F values exceeded those at the .05 level of the appropriate conservative distribution the selection of model or mastery level was deemed to affect different classification success levels.

In addition to learning if there were main effects, the study sought to determine if approaches differed. In order to determine if approaches as well as specific levels of significant factors differed, Scheffe's method of multiple comparisons was used. The contrasts of interest are given in Table 2.

Table 2. Contrasts for the model factor

	Contrast
1.	$(\bar{X}_{SIX} + \bar{X}_{TWELVE})/2 - (\bar{X}_{YHAT} + \bar{X}_{YHATP})/2$
2.	$(\bar{X}_{SIX} + \bar{X}_{TWELVE})/2 - (\bar{X}_{BAYES6} + \bar{X}_{BAYES12})/2$
3.	$(\bar{X}_{YHAT} + \bar{X}_{YHATP})/2 - (\bar{X}_{BAYES6} + \bar{X}_{BAYES12})/2$
4.	$\bar{X}_{SIX} - \bar{X}_{TWELVE}$
5.	$\bar{X}_{SIX} - \bar{X}_{YHAT}$
6.	$\bar{X}_{SIX} - \bar{X}_{YHATP}$
7.	$\bar{X}_{BAYES6} - \bar{X}_{BAYES12}$

CHAPTER IV

FINDINGS

This chapter contains the results of the analyses which were described in Chapter III. The findings are presented in five sections of which four are parallel to the discussion of the analysis and design in Chapter III. The first section gives the statistics which describe the variables of the study. Sections two through four will present the results of the three phases of analysis. The final section of this chapter will summarize the findings in terms of the three objectives which were stated in Chapter I.

Variables

Table 3 shows the descriptive statistics that were calculations for the twenty variables considered in this study. The statistics of the test variables of the study are given in Table 4. Since the domain test serves as the criterion for much of the analysis of this research, its reliability of .9777 is of particular importance.

Table 3. Descriptive statistics for all variables used in the study

Variable Name	Mean	Standard Deviation	Variance	Range
TEST	1.9277	.58745	.34509	1.0-4.0
TIME	9.2690	4.6209	21.352	2.5789-31.5
STEP	28.250	9.4406	89.124	8.0-48.0
SEX	1.4884	.50132	.25133	1. - 2.
SUBTEST(6)	5.0988	1.2219	1.4931	1. - 6.
SUBTEST(12)	9.5698	2.2271	4.9600	2. -12.
SUBTEST(18)	14.087	3.3558	11.261	1. -18.
SUBTEST(24)	18.174	5.1086	26.098	4. -24.
SUBTEST(30)	24.198	4.7765	22.814	5. -30.
SUBTEST(36)	28.831	5.5568	30.878	5. -36.
SUBTEST(42)	33.041	7.6917	59.162	2. -42.
SUBTEST(48)	37.174	8.7234	76.098	4. -47.
SUBTEST(54)	41.826	10.152	103.07	5. -54.
SUBTEST(60)	47.570	10.978	120.53	7. -60.
DOMAIN	122.53	28.531	814.03	14.-154.
SIX	.6213	.4857	.2359	0 - 1.
TWELVE	.7973	.4025	.1620	0 - 1.
VHAT	.6987	.4595	.2111	0 - 1.
VHATP	.6880	.4639	.2152	0 - 1.
BAYES6	.6853	.4648	.2162	0 - 1.
BAYES12	.7867	.3767	.1419	0 - 1.

Table 4. Descriptive statistics for DOMAIN, SUBTEST(J) and items comprising the six objectives

Test	Mean	Standard Deviation	Hoyt Reliability	Standard Error
DOMAIN	122.5291	28.5312	.9777	4.2435
OBJECTIVE 1	21.4128	4.2397	.8558	1.5784
OBJECTIVE 2	22.6337	4.2385	.8890	1.3850
OBJECTIVE 3	20.8837	5.3375	.8994	1.6604
OBJECTIVE 4	20.0233	6.4753	.9336	1.6362
OBJECTIVE 5	19.8488	5.4869	.8940	1.7516
OBJECTIVE 6	17.7267	6.2837	.9067	1.8829
SUBTEST(6)	5.0988	1.2219	.6158	.6914
SUBTEST(12)	9.5698	2.2271	.6900	1.1873
SUBTEST(18)	14.0872	3.3558	.8082	1.4284
SUBTEST(24)	18.1744	5.1086	.8843	1.7014
SUBTEST(30)	24.1977	4.7765	.8400	1.8786
SUBTEST(36)	28.8314	5.5568	.8699	1.9761
SUBTEST(42)	33.0407	7.6917	.9212	2.1328
SUBTEST(48)	37.1744	8.7234	.9259	2.3493
SUBTEST(54)	41.8256	10.1523	.9362	2.5399
SUBTEST(60)	47.5698	10.9785	.9423	2.6148

Phase 1

Based on very low correlations of SEX with the other information variables (see Table 5), it was eliminated from further consideration in the study.

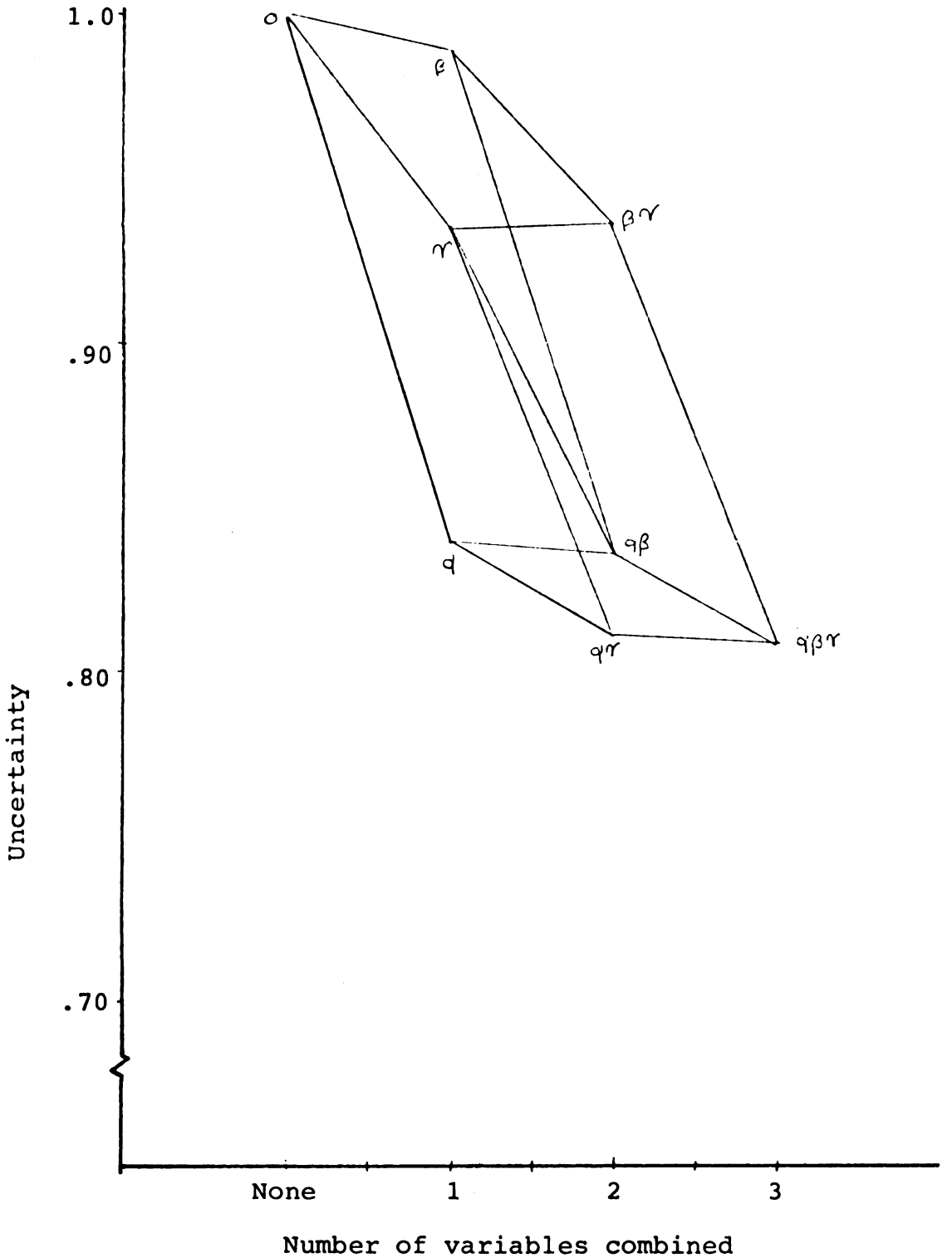
Figure 1 depicts graphically the information value of various combinations of the variables STEP, TEST, and TIME. The vertical axis represents the coefficient of alienation while the horizontal indicates the configuration of variables under consideration. By following the lines on the graph from left to right one can gain insight into the uncertainty reduction which will accrue by adding the indicated variable. If one compares the slopes of the line segments with the same initial point but different ending points, the relative informational value of the added variable will be apparent. For example, comparing the slope of $o\beta$ with $o\alpha$ indicates that STEP provides more information about the dependent variable than does TEST. In fact, examination of the segments representing the addition of TEST to equations, indicates that TEST contributes very little (if any) information. TIME appears to provide some information, but not as much as STEP. Another way of looking at the value of a variable's information is seen in Table 6.

Study of the sixth table confirms that STEP is

Table 5. Intercorrelation of information variables and domain achievements

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1.	SEX															
2.	STEP	-12*														
3.	TIME	-08	-27													
4.	TEST	08	-31	61												
5.	DOMAIN	02	55	-37	-26											
6.	SUBTEST(6)	-08	36	-26	-19	64										
7.	SUBTEST(12)	03	52	-37	-19	91	65									
8.	SUBTEST(18)	02	59	-36	-29	92	66	81								
9.	SUBTEST(24)	04	53	-34	-28	95	55	82	90							
10.	SUBTEST(30)	02	55	-38	-29	94	59	87	88	88						
11.	SUBTEST(36)	05	55	-34	-23	96	59	87	89	90	92					
12.	SUBTEST(42)	01	52	-34	-24	98	61	87	90	93	89	93				
13.	SUBTEST(48)	02	54	-38	-26	98	60	88	90	93	91	95	96			
14.	SUBTEST(54)	-01	53	-35	-27	98	58	89	90	95	94	95	96	97		
15.	SUBTEST(60)	01	54	-38	-29	99	62	88	90	94	93	94	96	97	97	

* all decimal points are omitted



α: STEP
β: TEST
γ: TIME

Figure 1. Reduction of uncertainty for combinations of 3 information variables

Table 6. Stepwise regression statistics for all permutations of the information variables with DOMAIN

Equation	Step	Variables	Multiple Correlation	Coefficient of Determination	Change in Coef. of Determination	p<
1	1	STEP	.5515	.3041	.3041	.000
1	2	TEST	.5590	.3125	.0084	.152
1	3	TIME	.5970	.3565	.0439	.001
2	1	STEP	.5515	.3041	.3041	.000
2	2	TIME	.5953	.3543	.0502	.000
2	3	TEST	.5970	.3565	.0021	.457
3	1	TIME	.3663	.1342	.1342	.000
3	2	TEST	.3690	.1362	.0020	.535
3	3	STEP	.5970	.3565	.2203	.000
4	1	TIME	.3663	.1342	.1342	.000
4	2	STEP	.5953	.3543	.2201	.000
4	3	TEST	.5970	.3565	.0021	.457
5	1	TEST	.2598	.0075	.0675	.001
5	2	TIME	.3690	.1362	.0687	.000
5	3	STEP	.5970	.3565	.2203	.000
6	1	TEST	.2598	.0675	.0675	.001
6	2	STEP	.5590	.3215	.2450	.000
6	3	TIME	.5970	.3565	.0439	.000

the most informative variable. Time is the second most informative. Examination of the results of the six equations (especially equation 4) suggests that TEST has no useful relationship with DOMAIN which is independent of STEP and TIME. Table 7 gives additional insight into the relationships of the information inherent in the four variables. In particular, the zero order partials for the three information variables indicate that each has a significant information factor relative to DOMAIN. However, the lack of significance of the first and second order partials $r_{D\beta.\alpha}$, $r_{D\beta.\gamma}$ and $r_{D\beta.\alpha\gamma}$ suggest that the information in TEST relative to DOMAIN is accounted for by STEP and TIME.

Based on the data presented in Tables 6 and 7 and the earlier elimination of sex as a useful variable, the most parsimonious regression equation relating non-test informational variables to DOMAIN included the independent variables TIME and STEP only. The basic statistics for this equation are presented in Table 8.

In order to determine the utility of including test information in the decision process regarding domain achievement, incremental validity was explored. Table 9 provides the data for assessing this incremental validity. The base, non-test, information accounts

Table 7. Partial correlations and coefficients of alienation for the information variables with DOMAIN

Zero Order	$r_{D\alpha}$= .551*	$r_{D\beta}$=-.260*	$r_{D\gamma}$=-.366*
	$K_{D\alpha}$= .835	$K_{D\beta}$= .966	$K_{D\gamma}$= .931
	$r_{\alpha\beta}$=-.313*	$r_{\alpha\gamma}$=-.273*	$r_{\beta\gamma}$= .613*
	$K_{\alpha\beta}$= .950	$K_{\alpha\gamma}$= .962	$K_{\beta\gamma}$= .790
First Order	$r_{D\beta.\alpha}$=-.110	$r_{D\gamma.\alpha}$=-.269*	$r_{\beta\gamma.\alpha}$= .577*
	K = .994	K = .963	K = .817
	$r_{D\alpha.\beta}$= .513*	$r_{D\gamma.\beta}$=-.271*	$r_{\alpha\gamma.\beta}$= .1081
	K = .858	K = .963	K = .9941
	$r_{D\alpha.\gamma}$= .504*	$r_{D\beta.\gamma}$= .048	$r_{\alpha\beta.\gamma}$= .192
	K = .864	K = .999	K = .981
Second Order	$r_{D\gamma.\alpha\beta}$=-.253*	$r_{D\beta.\alpha\gamma}$= .057	$r_{D\alpha.\beta\gamma}$= .505*
	K = .967	K = .998	K = .863

*Significant at .05 level

D: Domain

α: STEP

β: TEST

γ: TIME

Table 8. Regression statistics for relating TIME and STEP to DOMAIN

Variable	Standardized Regression Coefficient	T-Value	Significance Level
STEP	.4878	7.59077	.0000
TIME	-.2330	-3.62596	.0004

Table 9. Statistics for incremental validity analysis

Variables	Multiple Correlation	Coefficient of Determination	Incremental Validity	Coefficient of Alienation	F
Base (STEP, TIME)	.5953	.3543		.8036	*46.370
SUBTEST(6)	.7431	.5521	.1478	.6692	*74.207
SUBTEST(12)	.9120	.8317	.3167	.4102	*476.424
SUBTEST(18)	.9225	.8510	.3272	.3860	*559.810

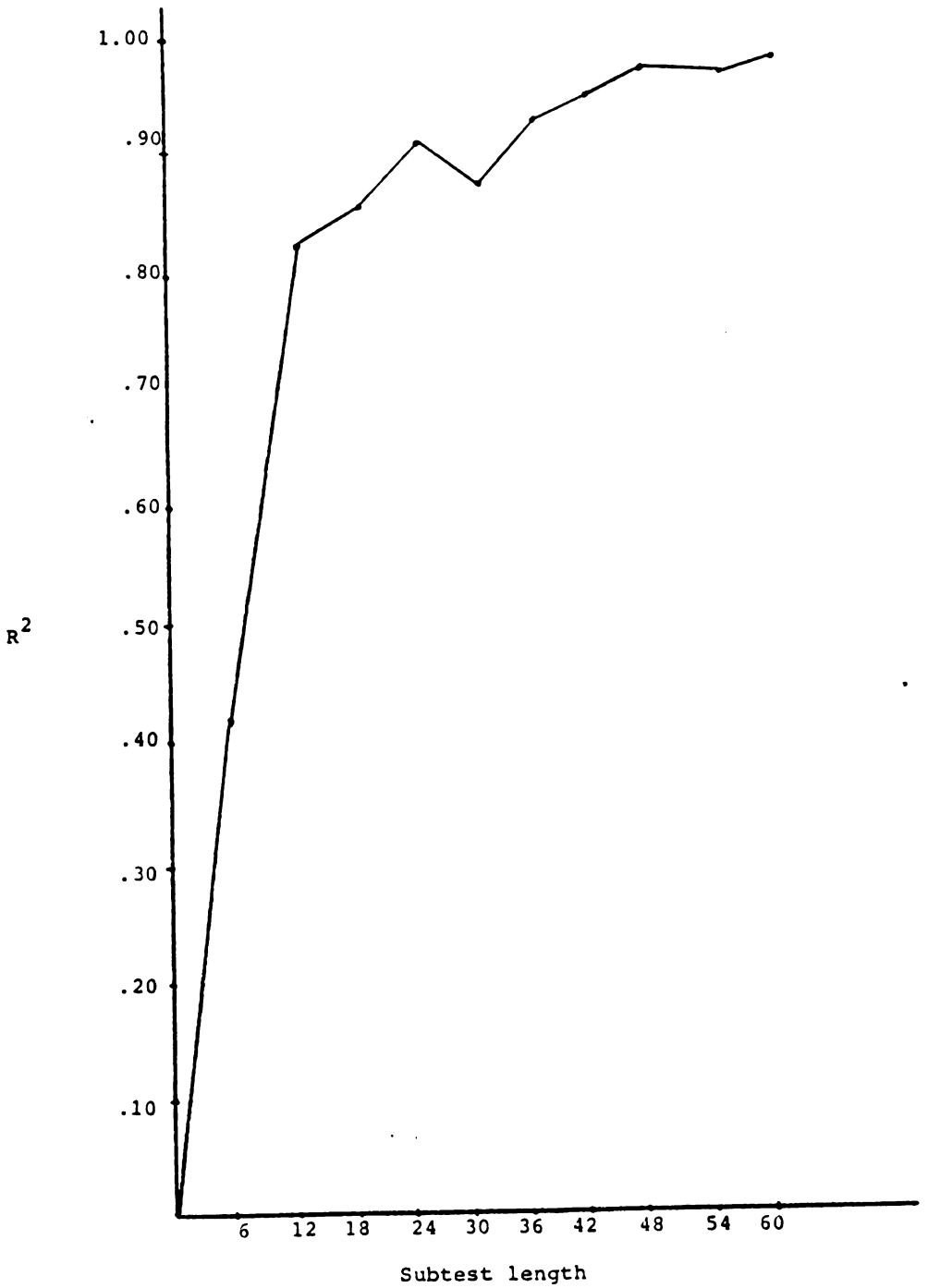
for over thirty-five percent of the variance in the dependent variable, domain achievement. The six item subtest result accounts for an additional twenty percent. If it is assumed that the relationship between information and test length is approximately linear in the interval between six and twelve items, a ten item test will augment the information in the base variables by an amount equal to that contained by the base variables. Since the F tests listed in Table 9 are for the partial regression coefficients relative to the dependent variable DOMAIN, each test significantly augments the base variables.

Tables 9 and 10 may be used to deduce the functional length of SUBTEST(6) coupled with STEP and TIME. Table 9 shows that the coefficient of determination (R^2) for the base variables plus SUBTEST(6) is .5521. Reference to Table 10 allows one to see that this R^2 value lies between the r^2 for SUBTEST(6) and SUBTEST(12). The graph of Figure 2 shows the relationship between the length of the subtests and the corresponding r^2 with domain. Based on this graph, it seems reasonable to obtain the functional length of the six item test by linear interpolation. This process yields a value of 8.03 which is the functional length of the 6 item subtest augmented by the two information variables. Clearly, the coefficients of determination in Tables

Table 10. Coefficients of correlation and determination for SUBTEST(J) with DOMAIN

Variables	Correlation Coefficients	Coefficients of Determination
SUBTEST(6)	.643	.413
SUBTEST(12)	.907	.823
SUBTEST(18)	.922	.850
SUBTEST(24)	.954	.910
SUBTEST(30)	.936	.876
SUBTEST(36)	.960	.920
SUBTEST(42)	.975	.940
SUBTEST(48)	.983	.966
SUBTEST(54)	.980	.960
SUBTEST(60)	.987	.974

Figure 2. The relationship of R^2 and subtest length



9 and 10 suggest that the antecedent variables provide a decreasing amount of information in combination with test data as the number of items in subtests of the domain increases. In effect, the functional lengths of subtests containing 12 or more items is the same as the length of the specific subtest.

Phase II

Tables 11 and 12 contain the basic statistics of the two least squares regression models which are to be the basis for classification. The first of these two tables contains only the information variables STEP and TIME. Table 12 presents the statistics for the information model with SUBTEST(6) added. The standard errors of these two models are 22.42 and 19.40 respectively.

The statistic of the Bayesian model which is roughly analogous to regression weights of the least square model is ρ^*. Table 13 presents values of ρ^* for three values of t which span the range of t values used in this study. Numbers are given for SUBTEST(6), SUBTEST(12), and SUBTEST(18). Reference to equation (22) of Chapter II suggests that as t increases, the influence of the mean becomes larger. Also, in all cases the influence of a subjects score becomes greater

Table 11. Regression statistics for TIME and STEP with DOMAIN

Variable	Standardized Regression Coefficient	T-Value	Significance Level
STEP	.5647	7.50541	.0000
TIME	-.3123	-4.15125	.0001

Table 12. Regression statistics for TIME, STEP and SUBTEST(6) with DOMAIN

Variable	Standardized Regression Coefficient	T-Value	Significance Level
STEP	.4238	6.08734	.0000
TIME	-.2124	-3.15086	.0022
SUBTEST(6)	.4094	5.70972	.0000

Table 13. ρ^* values for three values of t for subtests of length 6, 12 and 18

Test Size	t= 2.75	t= 9.5	t= 18
Six Item	.42697	.29335	.22126
Twelve Item	.57641	.48499	.43870
Eighteen Item	.67761	.617086	.58620

as test length increases. Figures 3, 4, 5, 6 and 7 illustrate the effect of the t parameter on the classification of the Bayesian Model. One can see that in all cases the most accurate classification can be achieved with t set equal to 2.75. Thus for the purpose of comparing models, it was judged appropriate to use the Bayesian Model with t equal to 2.75. This is the apparent "best" Bayesian model available for the present data.

The means and variances of the two raw score decision criterion are given previously in Table 1.

Phase III

The concluding set of findings yield information about the relative effectiveness of the three approaches for making decisions about mastery or non-mastery of the achievement domain. Table 14 presents the number and percentage of correct classifications for each of the six models at each mastery level. The remainder of this section discusses results of the statistical analysis of these data.

As is shown in Table 15, the analysis of variance yielded significant mastery level and model effects. With respect to the mastery level factor, the proportion of correct classifications appears to decrease as mastery levels increases. This can be seen in Table 16.

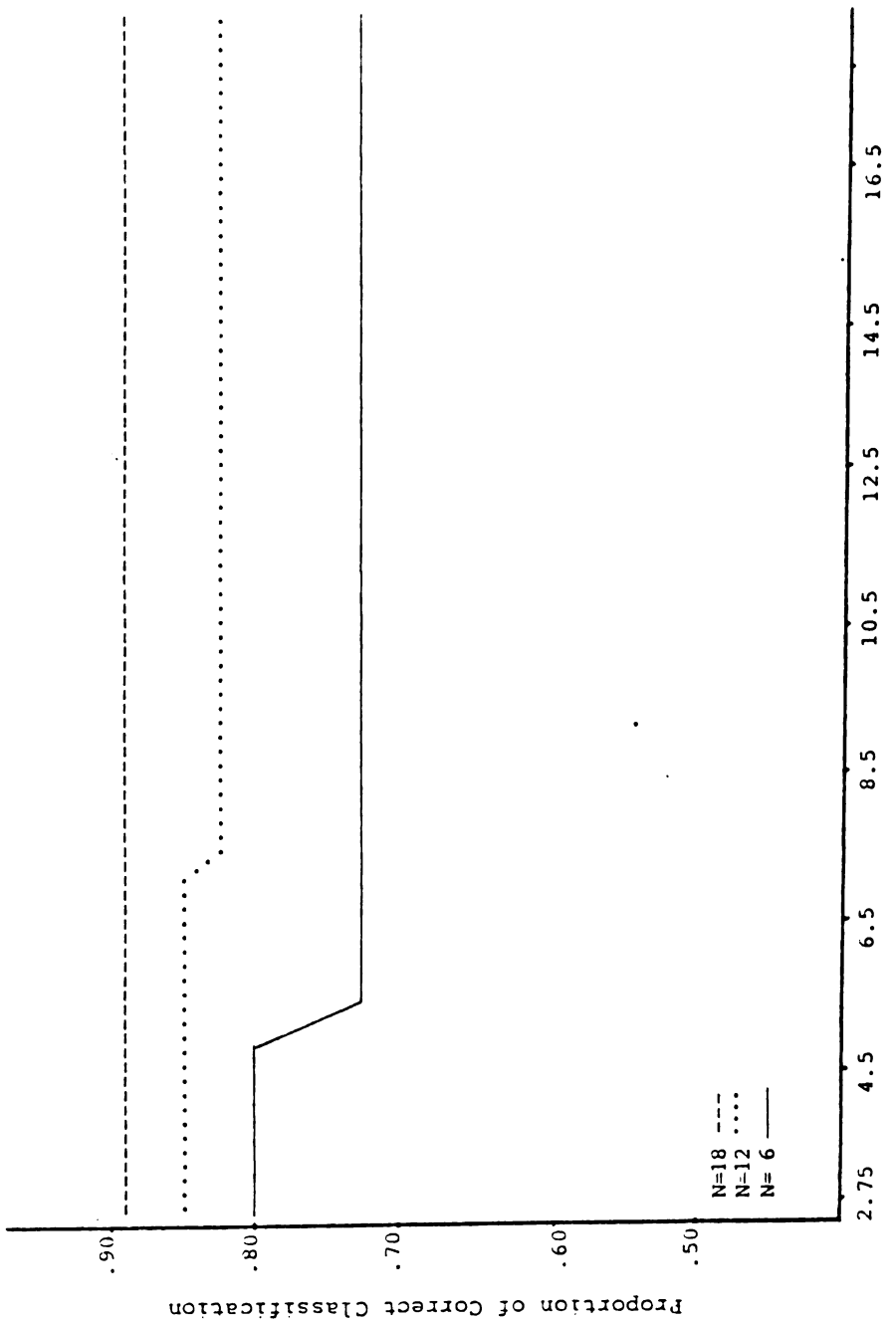


Figure 3. The relationship of t to correct classification for the master level of .70

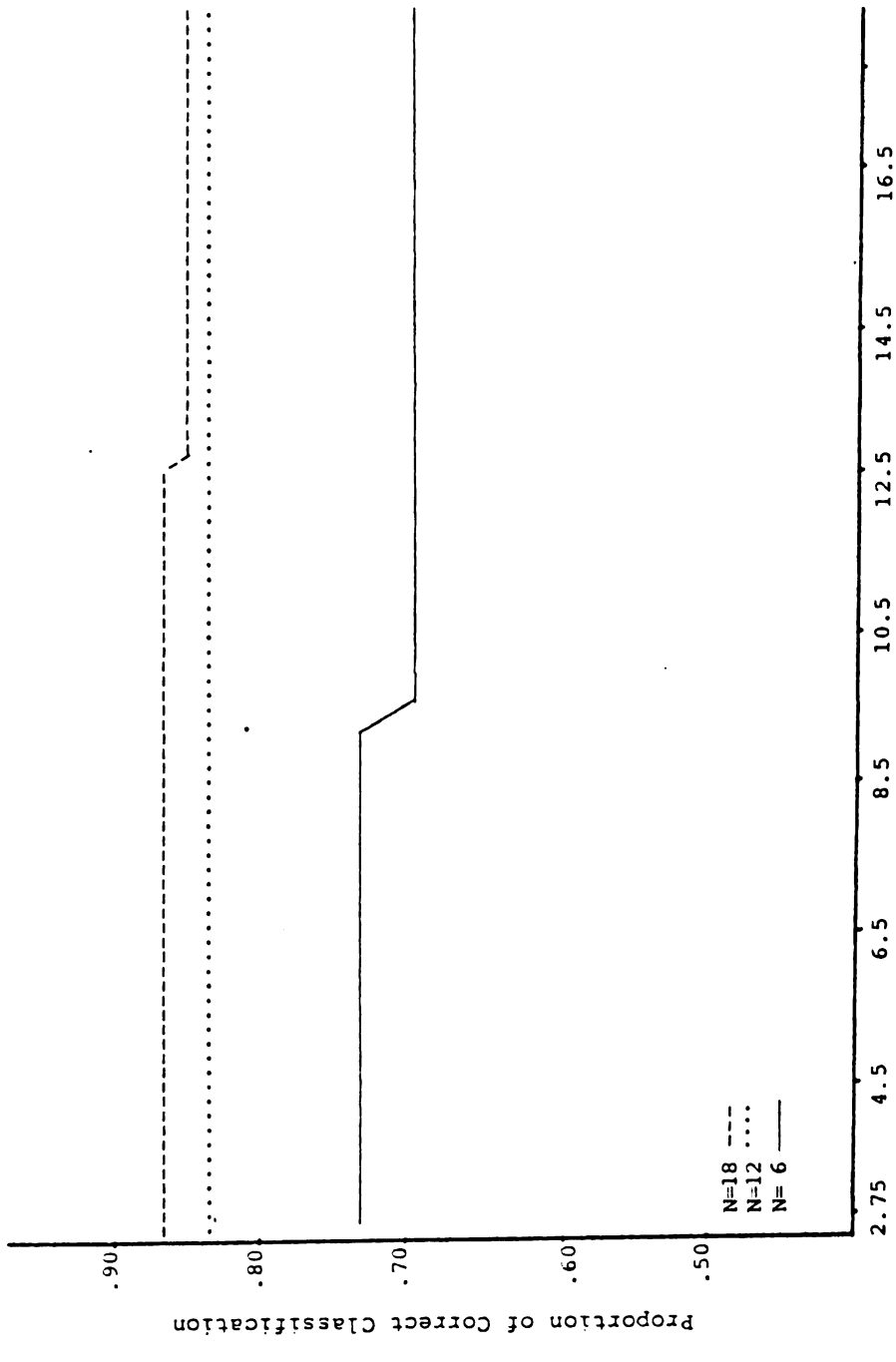


Figure 4. The relationship of t to correct classification for the master level of .75

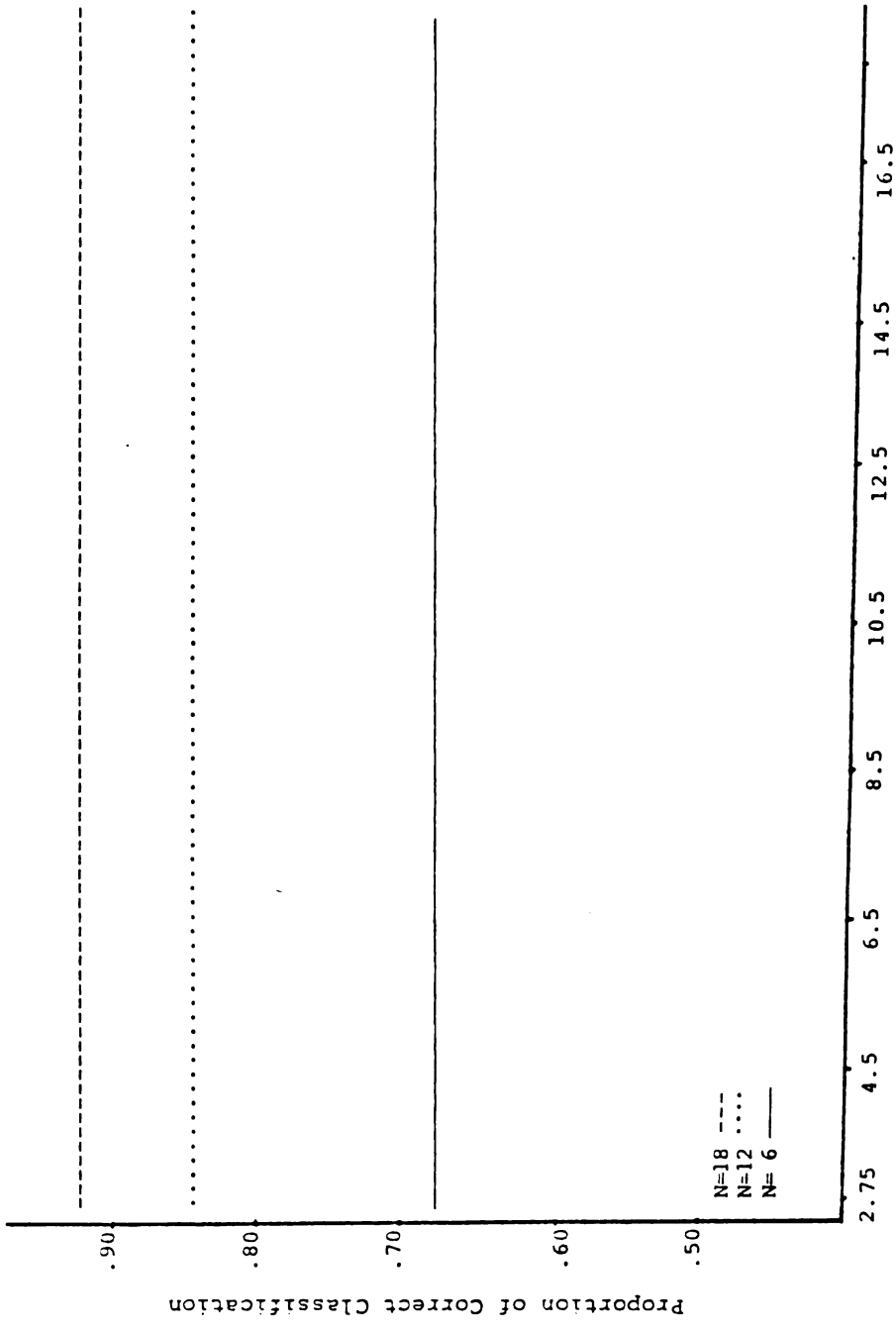


Figure 5. The relationship of t to correct classification for the master level of .80

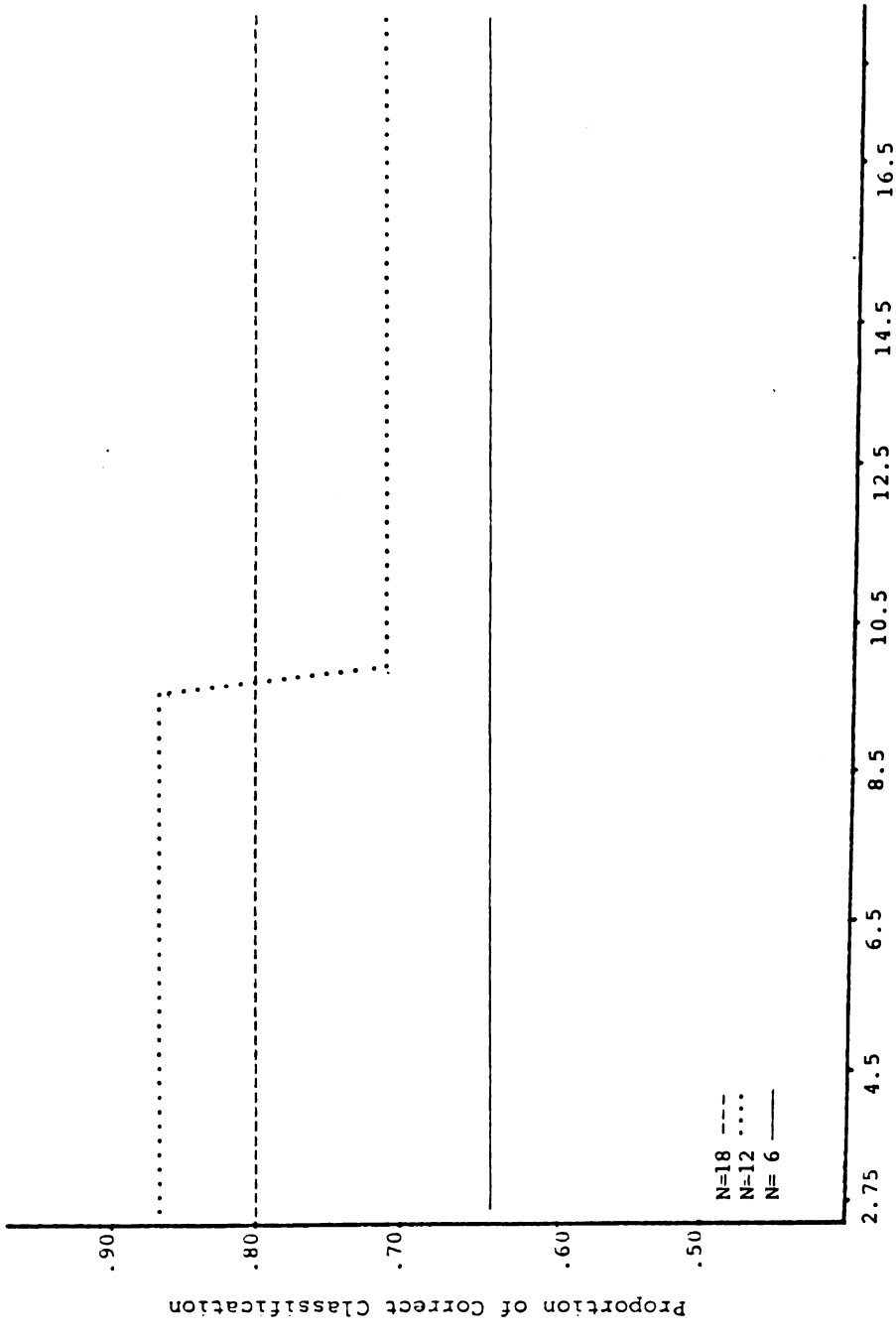


Figure 6. The relationship of t to correct classification for the master level of .85

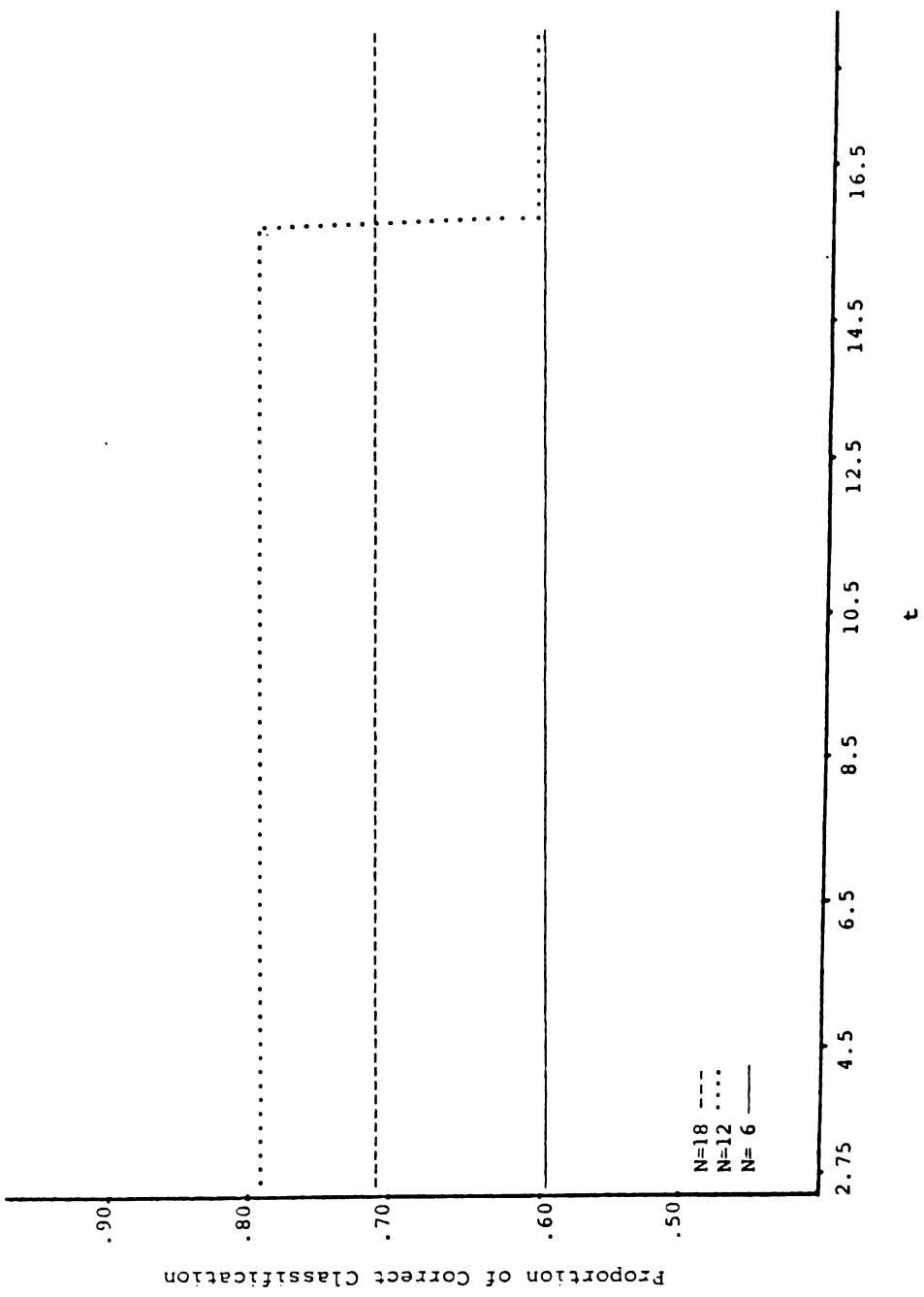


Figure 7. The relationship of t to correct classification for the master level of .90

Table 14. Number and percent of correct classifications

Mastery Level		Six	Twelve	Model YHAT	YHATP	BAYES6	BAYES12
70	%	84.0	85.3	77.3	78.7	80.0	85.3
	N	63	64	58	59	60	64
75	%	73.3	82.7	73.3	72.0	73.3	82.7
	N	55	62	55	54	55	62
80	%	66.7	82.7	65.3	65.3	66.7	82.7
	N	50	62	4	49	50	62
85	%	60.0	81.3	64.0	62.7	62.7	85.3
	N	45	61	48	47	47	64
90	%	54.7	82.7	73.3	68.0	60.0	78.7
	N	41	62	55	51	45	59

Table 15. Analysis of variance statistics

Source	df	MS	F	Conservative F Distribution
Model	5	1.95804	10.40*	$_{.05}F_{1,4} = 7.71$
Mastery Level	4	1.26511	6.72*	$_{.05}F_{1,5} = 6.61$
Interaction	20	.16671	.89	
Within	2220	.18831		*Significant at .05 level

Table 16. Means and variances for levels of model and mastery level

Factor	Level	Mean	Variance
Model	SIX	.6773	.2191
	TWELVE	.8293	.1419
	YHAT	.7067	.2078
	YHATP	.6933	.2132
	BAYES6	.6853	.2162
	BAYES12	.8293	.1419
Mastery Level	.70	.8178	.1493
	.75	.7622	.1816
	.80	.7156	.2040
	.85	.6933	.2131
	.90	.6956	.2122

The Scheffe' contrasts in Table 17 suggests that the model effect stems from the classification differences between the two raw score models and the two Bayesian models. There are no significant differences between the three decision approaches. It is notable that the variances for correct classification by SUBTEST(12) and BAYES12 is considerably lower than is the case for the other four models. This fact is appropriately considered in consort with change in R^2 values between SUBTEST(6) and SUBTEST(12) in Table 6.

This chapter has summarized the findings of the three phases of analysis. Initially, the utility of the information variables for reducing uncertainty about domain achievement was reported. Then the parameters of the decision models were presented. Finally, the results of the statistical comparisons of the six models were given. The final chapter of this thesis discusses the implications of the findings and presents conclusions which can be drawn from them.

Table 17. Scheffe' contrast statistics for the model factor

Contrast ($\hat{\Psi}$)	Variance of Contrast $\sigma^2_{\hat{\Psi}}$	$\Psi/\hat{\sigma}_{\hat{\Psi}}$
$\frac{\bar{X}_{SIX} + \bar{X}_{TWELVE}}{2} - \frac{\bar{X}_{YHAT} + \bar{X}_{YHATP}}{2}$	.0022	1.1304
$\frac{\bar{X}_{SIX} + \bar{X}_{TWELVE}}{2} - \frac{\bar{X}_{BAYES6} + \bar{X}_{BAYES12}}{2}$	.0022	- .08483
$\frac{\bar{X}_{BAYES6} + \bar{X}_{BAYES12}}{2} - \frac{\bar{X}_{YHAT} + \bar{X}_{YHATP}}{2}$	.0022	-1.2153
$\bar{X}_{SIX} - \bar{X}_{TWELVE}$	.0272	-5.588*
$\bar{X}_{SIX} - \bar{X}_{YHAT}$	.0272	-1.081
$\bar{X}_{SIX} - \bar{X}_{YHATP}$	.0272	- .588
$\bar{X}_{BAYES6} - \bar{X}_{BAYES12}$	.0272	-5.294*
$\bar{X}_{SIX} - \bar{X}_{BAYES6}$	.0272	- .293

* $\Psi/\hat{\sigma}_{\hat{\Psi}} > {}_{.05}F_{1,4} = 7.71$

CHAPTER V

INTERPRETATION, CONCLUSIONS, RECOMMENDATIONS

In this chapter the data presented in the previous chapter are evaluated in terms of the objectives given in Chapter I. Conclusions based on the evaluation are also given along with recommendations for practice and subsequent research.

The first objective of this project was:

> to determine the information existent in four antecedent and collateral variables relative to domain achievement.

It should be recalled that data are considered information if and only if it reduces the uncertainty involved in making a decision. Analysis of the four information variables suggested that only two truly yielded information. Sex was unrelated to any of the variables of the study. TEST, while correlated with domain achievement, contained no information not present in TIME. The other variable which contained information relative to DOMAIN was STEP achievement.

The two significant information variables indicate prior mathematics achievement and learning rate. The relationship between prior mathematics achievement and subsequent test performance was certainly expected.

The fact that learning rate had predictive utility independent of achievement is of interest. This finding is consistent with Carroll's (1963) hypothesis that time is a central factor in achievement. The findings of this research suggest that if two pupils have identical prior achievement and different prior learning rates, the student with the higher rate will be expected to score higher on subsequent achievement measures. Thus, in terms of estimating posterior scores everything else being equal, quicker students should surpass the less quick ones. In addition, students with slightly inferior achievement but higher learning rates should be expected to catch pupils with higher achievement but lower learning rates. It seems reasonable to conclude that in the long run if opportunity and motivation are equal the advantage will always be with the quicker student.

The classroom teachers trying to summarize the useful prior information they possess relevant to subsequent achievement should consider both achievement and rate of learning. Achievement level seems to be most important; however, rate, being a dynamic variable, should be considered in terms of the length of time which has passed since the last appraisal of achievement level. Gettinger and White (1979) have recently reported

an approach to measuring time to learn which would allow teachers in traditional classroom settings to easily appraise learning rate. It is recommended that teachers familiarize themselves with their approach and apply it routinely.

The procedure is as follows: pupils study standard materials, which they have not mastered, for a specified length of time and are then tested. This is repeated until mastery at some arbitrary upper limit has been reached. Time to learn is then said to be the number of trials required. The cited authors had students follow the process for six types of tasks and set time to learn as the mean number of trials needed for mastery.

The second objective of this study aimed to determine:

1) the incremental validity of short domain tests,
2) if decision precision can be improved by using antecedent and collateral data with test results, and
3) the functional lengths of several short domain tests.

Incremental validity refers to the extent to which a multiple correlation is raised by the addition of test results to a set of prior existing information.

Thus the incremental validity of SUBTEST(6), SUBTEST(12), and SUBTEST(18) is .1478, .3167, and .3272 respectively. The incremental validity of the six item test is less one quarter of the base information (assuming no prior information with respect to the bases). Cronbach and Gleser (1965) have written that "tests should be judged by the increase in validity which they offer." In terms of information, as this study has defined it, the six item test does provide some. In order to determine if the amount of information is meaningful with respect to mastery-nonmastery decisions, the decision precision based on the prior information and the prior information combined with the six items subtest was compared. (It should be recalled that "decision precision" has previously been defined as the proportion of correct classifications made on the basis of a given decision algorithm. The decision based on the application of the algorithm to the domain achievement score is the correct one.)

The results of this comparison indicated that decision precision was not improved by using the six item test. The implication of this finding is clear. Test data do not provide decision relevant information that was not available prior to testing. Thus, while use of the tests might be justified on instructional

grounds, a decision to test with six items is not justifiable as a means of improving decisions about mastery-nonmastery. This is true regardless of whether the prior information is incorporated by least squares or Bayesian approach.

The number of test items necessary to provide information equivalent to that of the collateral, antecedent and test information already available is referred to as the functional length of a test. Thus TEST, TIME and SUBTEST(6) have a functional length of 8.03. One could use Figure 2 to set a functional length for the base prior information. The value would be slightly more than five. It is clear that if one considers the prior information and then the six item test, the information value of the test is reduced to that of about three items. The findings of phase III of the analysis suggest that this is not a sufficient number of items to improve decision precision, vis-a-vis mastery-nonmastery, significantly.

For subtests of 12 items or more the functional length is the same as the actual length. Thus one would expect that the decision precision of an algorithm incorporating prior information would be the same as one based solely on test score.

To address the final objective of this research,

comparisons of the three decision approaches were made. With respect to decision precision the three approaches do not differ.

In order to spur insights into the result of no difference among the approaches it is useful to compare the approaches in detail. While each of the models is linear, the least squares approach is not directly comparable algebraically to the other two. However, the Bayesian and raw score approach are analogous and comparison of their algebraic basis is instructive.

In order to do this, one should recall the Kelley model for estimating true scores. The Kelley model is

$$\hat{T} = \rho_{xx'} X + (1-\rho_{xx'})\overline{X}$$

Where $\rho_{xx'}$ is the proportion of true to observed variance, X is an observed score and $\overline{X}$ is the mean of such scores ($\overline{T} = \overline{X}$). The raw score approach is the specific case where $\rho_{xx'} = 1$ and thus $\hat{T} = X$.

The Bayesian Marginal Mean Model has the same form as Kelley's Model. Like Kelley's approach, it contains a parameter which is, in part, a function of score variance. However, this parameter is also influenced by prior subjective estimates about the sample in question. Specifically, this prior information

is incorporated into the model by specification of a value for prior information in terms of the number of test items the information is worth. Table 13 indicates that ρ^* is clearly a function of t. However, Figures 3 through 7 as well as the results of the Scheffe' contrasts suggest that the decision about mastery-nonmastery is not particularly sensitive to t. It appears that for the purpose of classifications of mastery or non-mastery, incorporation of prior information by means of t has little value. For after the complex calculations of the Bayesian Model are completed it functions as the raw score form of Kelley's Model.

For making the kinds of decisions made most frequently by educators, the raw score model is clearly indicated because of its simplicity.

The following three points summarize the comparison of the models.

1. Decision precision was the same for the six item raw score model and the least square model containing only antecedent and concommitant information.
2. Decision precision was improved when 12 items were used rather than six.
3. The raw score model is preferred to the Bayesian model.

All of the preceding discussion holds for mastery levels of .70, .75, .80, .85, and .90. However, across all models the precision decreases as the mastery level is increased. This trend does not appear to be uniform for all models. It seems as though the models containing the least information decline most in precision. Both the raw and Bayesian approaches using six items show the greatest consistent decline.

The finding of this research which seems to have the greatest utility for current classroom practice is that selected prior information appropriately weighted, can be used to yield decisions about subsequent achievement which are as accurate as decisions based on a six item test. This fact can be useful as teachers informally monitor pupils on a day to day or even minute to minute basis. Assuming that a teacher has prior measures of achievement and rate of learning is invariant (at least within a subject and group of pupils) it may be sufficient to keep track only of students on task behavior to assure they are progressing. Perhaps students can be taught that frustration in learning attempts signals a diagnosis point where they should ask for help. If the teacher can't easily identify the problem, then a test of sufficient length to diagnose the difficulty is called for. It may be that the frequent tests called for by current

individualized programs are unnecessary. What may be called for instead is a sound initial placement of instructional materials and methods based on learning rate. After this start subsequent testing can be done when frustration is indicated by off task behavior or identified by the student.

Such an approach would probably result in some students taking frequent tests and others taking very few. It would reduce unnecessary assessment and assure that when a test was given its purpose would be clear to both teacher and student. Hopefully, it would allow tests with sufficient items to assure infrequent errors in instructional decisions. These suggestsion will need further investigation.

This research cannot be generalized beyond the curriculum and grade level of focus. Such extension would require further research. It is suggested that efforts be focused on issues related to classroom practice as discussed in the previous paragraphs rather than the replication of the present study.

LIST OF REFERENCES

LIST OF REFERENCES

Anscombe, F.J. The Transformation of Poisson, Binomial and Negative Binomial Data. Biometrika, 1948, 35, 246-254.

Baker, E.L. Beyond objectives: Domain-referenced tests for evaluation and instructional improvement. EDUC. TECH., 1974, 14, 10-16.

Bormuth, J.R. On the Theory of Achievement Tests, Chicago: University of Chicago Press, 1970.

Carroll, J.B. A model for school learning. Teachers College Record, 1963, 64, 723-733.

Cohen, J. and Cohen, P. Applied Multiple Regression/ Correlation Analysis for the Behavioral Sciences. Hilldale, New Jersey: Lawrence Erlbaum Associates, 1975.

Cronbach, L.J. and Gleser, G.C. Psychological tests and Personnel Decisions. University of Illinois Press, Urbana, 1965.

Cronbach, L.J. Test Validation In R.C. Thorndike (ed.) Educational Measurement: Second Edition, Washington, D.C.: American Council on Education, 1971.

Draper, N.R. and Smith, H. Applied Regression Analysis. New York: John Wiley & Sons, Inc., 1966.

Ebel, R.L. Criterion referenced measurements: limitations. School Review, 1971, 69, 282-288.

Fhaner, S. Item sampling and decision making in achievement testing. British Journal of Statistical Psychology, 1974, 27, 172-176.

Gettinger, M. and White, M.A. Which is the strongest correlates of school learning? time to learn or measured intelligence: Journal of Educational Psychology, 1979, 71, 405-412.

Glaser, R., and Nitko, A.J. Measurement in learning and instruction. In R.L. Thorndike (ed.) Educational Measurement. Washington: American Council on Education, 1971.

Greenhouse, S.W. and Geisser, S. On methods in the analysis of profile data. Psychometrika, 1959, 24, 95-112.

Harris, C.W., Alkin, M.C. and Popham, W.J. Problems in Criterion-referenced Measurement, CSE monograph series in evaluation. No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Harris, M.L. and Steward, D.M. Application of classical strategies to criterion referenced test construction. A paper presented at the Annual Meeting of the American Educational Research Association, 1971.

Hambleton, R.R. Testing and decision making procedures for selected individualized instructional programs. Review of Educational Research, 1974, 44, 371-400.

Hambleton, R.R., Novick, M.R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Hambleton, R.R., Swaminathan, H., Algina, J., and Coulson, D.B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Hively, W., Patherson, H.L., and Page, S.A. A "universe-defined" system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 5, 275-290.

Ivens, S.H. An investigation of item analysis, reliability and validity in relation to criterion-referenced tests. Unpublished Doctoral Dissertation. Florida State University, 1972.

Jackson, P.H. Simple approximations in the estimation of many parameters. British Journal of Mathematical and Statistical Psychology, 1972, 25, 213-229.

Kelley, T.L. Interpretation of Educational Measurements. Yonkers on Hudson, New York: World Book, 1927.

Kerlinger, F.N. and Pedhazur, E.J. Multiple Regression in Behavioral Research. New York: Holt, Rinehart and Winston, Inc., 1973.

Lewis, C., Wang, M.M. and Novick, M.R. Marginal distributions for the estimation of proportions in m groups. ACT Technical Bulletin No. 13. Iowa City, Iowa: The American College Testing Program, 1973.

Lord, F.M. and Novick, M.R. Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1968.

Mager, R.F. Preparing Instructional Objectives. Palo Alto, California: Fearson Publishers, Inc., 1962.

Millman, J. Determining test length: Passing scores and test lengths for objective-based tests. Instructional Objectives Exchange, Los Angeles, California, 1972.

Millman, J. Criterion-referenced measurement. In W.J. Popham (ed.) Evaluation in Education: Current Applications. Berkeley, California: McCutchan Publishing Co., 1974.

Millman, J. Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 1973, 43, 205-216.

Novick, M.R. and Jackson, P.H. Statistical Methods for Educational and Psychological Research. New York: McGraw-Hill, 1974.

Novick, M.R., Lewis, C. and Jackson, P.H. The estimation of proportions for M groups. Psychometrika, 1973, 38, 19-45.

Novick, M.R. and Lewis, C. Prescribing test length for criterion-referenced measurement. In C.W. Harris, M.C. Alkin and W.J. Popham (eds.) Problems in Criterion-Referenced Measurement. Monograph Series in Evaluation No. 3. Los Angeles, Center for the Study of Evaluation, University of California, 1974.

Osborn, H.G. Item sampling for achievement testing. Ed. and Psych. News, 1968, 28, 95-104.

Popham, W.J. Educational Evaluation, Englewood Cliffs, New Jersey: Prentice Hall, 1975.

Rozeboom, W.W. Foundations of the Theory of Prediction. Homewood, Illinois: The Dorsey Press, 1966.

Sechrest, L. Incremental validity: A recommendation. Educational and Psychological Measurement, 1963, 23, 153-158.

Swaminathan, H., Hambleton, R.R., Algina, J. A Bayesian decision-theoretic procedure for use with criterion-referenced tests. Journal of Educational Measurement, 1975, 12, 87-98.

Wang, M.M. Tables of constants for the posterior marginal estimates of proportions in M groups. ACT Technical Bulletin No. 14. Iowa City, Iowa: The American College Testing Program, 1973.

LIST OF NOTES

LIST OF NOTES

1. Hambleton, R.R., Swaminathan, H., Algina J., and Coulson, D. Criterion Referenced Testing and Measurement: A Review of Technical Issues and Developments. Unpublished Manuscript, University of Massachusetts, 1975.