DOCUMENT RESUME

ED 041 937 TM 000 029

AUTHOR	Pinsky, Paul D.
TITLE	A Mathematical Model of Testing for Instructional Management Purposes.
SPONS AGENCY	Charles F. Kettering Foundation, Dayton, Ohio.
PUB DATE	Mar 70
NOTE	19p.; From symposium "Designing Instructional Systems with Longitudinal Testing Using Item Sampling Techniques." (Annual meeting, American Educational Research Association, Minneapolis, Minn., March 1970)
EDRS PRICE	EDRS Price MF-$0.25 HC-$1.05
DESCRIPTORS	Behavioral Objectives, Instructional Improvement, Management, *Mathematical Models, Statistical Analysis, Symposia, *Test Construction, *Testing
IDENTIFIERS	*Comprehensive Achievement Monitoring

ABSTRACT

Developing a student testing mathematical model for instructional management purposes necessitates clear structuring of the curriculum materials involved, whether designated in the domain of content or the dimension of concepts or skills. Such structuring of a course written in performance objectives is presented and noted to be helpful in making decisions concerning the construction of tests over time and in understanding the inter-relationships of the parts of the curriculum from the testing results. An algorithm to select test items used to estimate desired parameters is developed. Inputs into the model are the average time required to answer each item, the errors of measurement associated with each item, the relative value of the information provided, the prior knowledge of this information, and a value function on the accuracy of the resultant estimates. Finally, techniques are given to allocate test items to students in such a way as to generate simultaneous estimates of item and student group characteristics. References are included. (Author/ES)

A Mathematical Model of Testing for Instructional Management Purposes[1,2]

Paul D. Pinsky
Stanford University

1 A paper presented at the Annual Meeting of the American Educational Research Association, Minneapolis, March, 1970 in a symposium entitled "Designing Instructional Systems with Longitudinal Testing Using Item Sampling Techniques".

2 The research and development reported herein was performed pursuant to a grant from the Charles F. Kettering Foundation to the principal investigator, Dwight W. Allen, and under the project director, William P. Gorth, both of the School of Education, The University of Massachusetts, Amherst, Massachusetts 01002.

Considering a course written in performance objectives, the paper explores the information needed by a teacher for instructional management purposes. Then, based upon these needs at a fixed point in time, an algorithm is developed to select the test items to be used to estimate the desired parameters. Inputs into the model are the average time required to answer each item, the errors of measurement associated with each item, the relative value of the information provided, the prior knowledge of this information, and a value function on the accuracy of the resultant estimates. Finally, techniques are given to allocate test items to students in such a way as to generate simultaneous estimates of item and student group characteristics.

1.0 Introduction

Consider a teacher who wishes to impart some knowledge to a group of students. This teacher has a brother-in-law who is an operations researcher for a management information firm. At a family summer outing, the teacher overheard his brother-in-law talking about information feedback, control theory, decision analysis, and a whole raft of unfamiliar jargon. As the teacher was on vacation for the summer, he decided to find out about all these powerful techniques and apply them to his teaching in the fall. After all, if his brother-in-law could increase efficiency by 100% or more in the business and industrial world, why couldn't he do the same in the classroom? Think, only 30 minutes a day for each class instead of the usual 60! After a few years of practice, he could write a book, form his own company, and retire with the knowledge that he had done his share to save mankind.

In order to develop a model of testing students for instructional management purposes, it is necessary to explore the types of information that would be useful to a teacher in a classroom environment. First of all, the structure of the material the instructor is attempting to teach must be clearly defined. This material will be viewed in the content domain in a structure as shown in figure 1 for the model presented herein. Most courses currently defined in performance objectives do not recognize the concept of measurable objectives (MO's). However, in a course containing 100 or more performance objectives, it is not feasible to obtain statistically reliable estimates of achievement levels in all or most of the objectives at the same time. This phenomenon (the bandwidth-fidelity dilemma) is explained in more detail in Cronbach [3], Chapter 8. Therefore the course in figure 1 has been

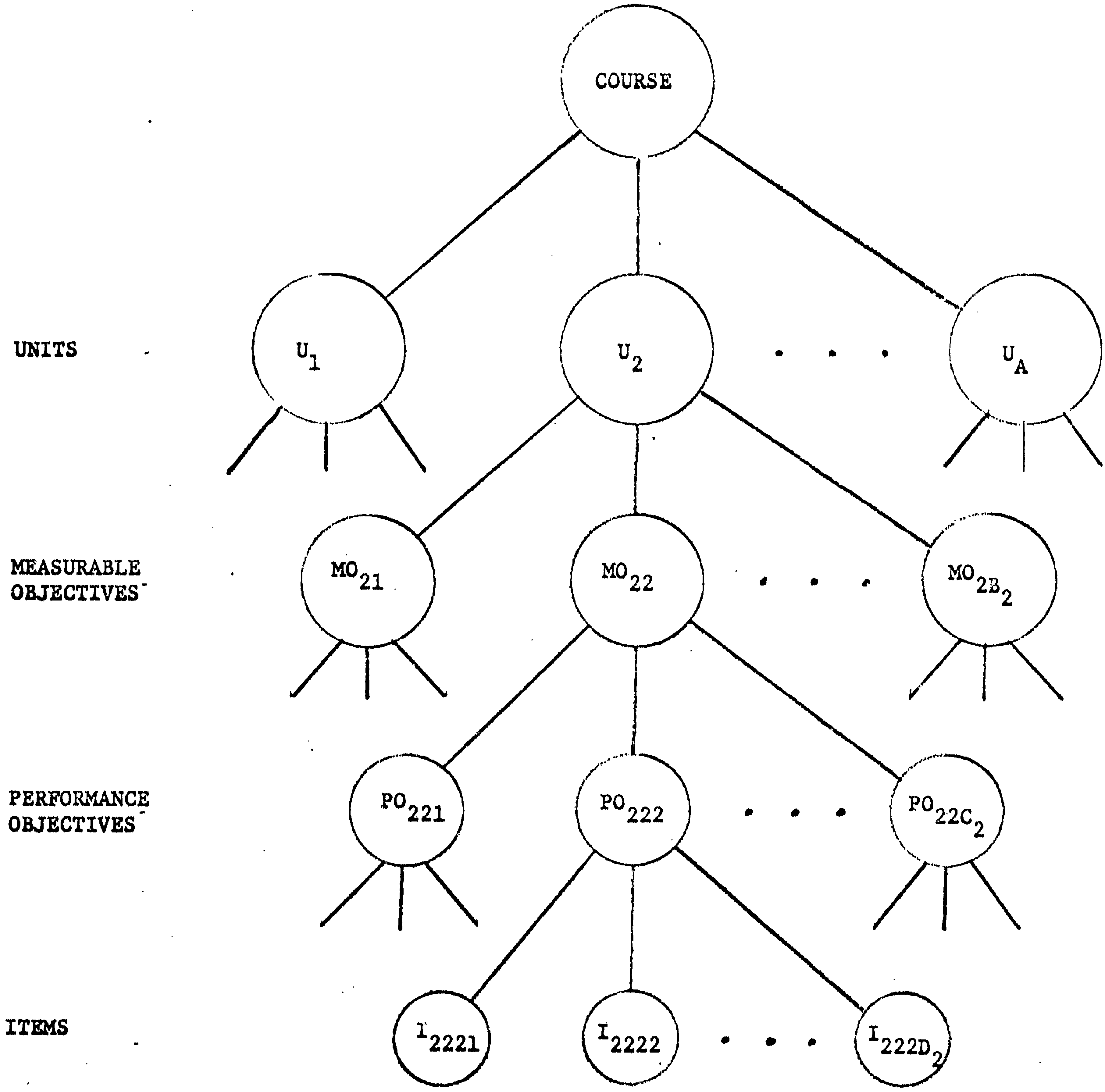
COURSE
UNITS
U_1
U_2
U_A
MEASURABLE OBJECTIVES
MO_{21}
MO_{22}
MO_{2B_2}
PERFORMANCE OBJECTIVES
PO_{221}
PO_{222}
PO_{22C_2}
ITEMS
I_{2221}
I_{2222}
I_{222D_2}

Figure 1

organized into groups of performance objectives large enough to yield reliable statistical estimates, but small enough to yield information valuable for instructional management purposes. CAM is presently using the performance objectives covered in a week's time as the set of measurable objectives. As will become clear later on, this concept of measurable objectives does not cause any loss of information concerning the individual performance objectives. This is important to note because data relating to performance objectives, although not statistically reliable, can still be quite meaningful to the teacher and student.

For notational purposes, the units will be indexed by a $\{U_a, a = 1,\cdots,A\}$, and measurable objectives by a (the unit they are in) and b $\{MO_{ab}, a = 1,\cdots,A, b = 1,\cdots,B_a\}$. Note that there can be different numbers of MO's per unit. Viewing the course in the structure as presented above is helpful for making decisions concerning the construction of tests over time and in understanding the inter-relationships of the parts of the curriculum from the results obtained by testing. Furthermore, it is assumed that the teacher has a presentation strategy; that is, a plan of the order in which the material will be presented to the class.

One can also view the curriculum in a second dimension, that of concepts or skills. An example of this dimension is the taxonomy developed by Bloom [1]. However, because of the difficulty in applying such classification schemes to curriculum and the increase in mathematical complexity in the present formulation, only the single dimensional classification by content will be considered here.

In order to discuss the possible goals of teaching a curriculum, some parameters must be defined. Let P_{abs} be the achievement level of student s in MO_{ab}. A convenient definition of P_{abs} is the expected percentage of correct responses by students to the universe of items that measures achievement in MO_{ab}. The estimate of P_{abs} can then be obtained by sampling items from the universal pool. $P_{ab.} \equiv \frac{1}{S} \sum_{s=1}^{S} P_{abs}$ is the average achievement level of the whole class in MO_{ab}. Now consider the goal of the teacher as maximizing $F(\{P_{abs}, a = 1,\cdots,A, b = 1,\cdots,B_a, s = 1,\cdots,S\})$. The following are specific examples of $F(\cdot)$ that teachers may choose:

$$F(\cdot) = \sum_{a=1}^{A} \sum_{b=1}^{B_a} P_{ab.}$$

This function values learning in all MO's and by all students equally.

$$F(\cdot) = \sum_{a=1}^{A} \sum_{b=1}^{B_a} \alpha_{ab} P_{ab.}$$

This function values learning by all students equally, but values learning in MC_{ab} according to the weight α_{ab}.

$$F(\cdot) = \sum_{a=1}^{A} \max_{b=1,\cdots,B_a} P_{ab.}$$

This function values learning one and only one MO within each unit very well.

$$F(\cdot) = \sum_{a=1}^{A} \min_{b=1,\cdots,B_a} P_{ab.}$$

This function values learning at least something about every MO in each unit.

The above are but a few examples of possible $F(\cdot)$. Others could differentiate between students, be non-linear functions of P_{abs}, or compare $P_{ab.}$ at the end with $P_{ab.}$ at the beginning of the course.

This digression into objective functions is presented to emphasize that a teacher should have a specific idea of his goals in order to make intelligent decisions regarding instructional strategies. In the next section, an algorithm is presented for the selection of test-items to be given to students at a fixed point in time. This algorithm requires the teacher to input the value of certain information to him. Without a clearly defined set of goals, the teacher may have a difficult time selecting the appropriate values.

The reason that one needs an algorithm to select items for testing purposes is that there exists only a finite amount of time for such testing. In this paper, testing is considered as being done at fixed intervals (called periods---say once a week) for a fixed amount of time each period (T---say 30 minutes). Varying the frequency and length of testing during the school year is not discussed, but the present model does not exclude such possibilities.

Finally consider the types of information that a teacher might like to have for each MO_{ab}. The value of $P_{abs}, s = 1,\cdots,S$ before instruction on MO_{ab} would certainly be useful. So would the values immediately after instruction and several months after instruction (the retention level). One of the problems is that P_{abs} is very difficult to estimate in a statistically reliable fashion due to sampling and measurement errors. $P_{ab.}$ is easier to measure and is the quantity that will be estimated. $P_{ab.}$ as defined contains information about all the students in the course. Section 4.0 explains how estimates of $\sum_{S \epsilon U_g} P_{abs}, g = 1,\cdots,G$ can be simultaneously generated where U_g is a subset of all students. Examples of U_g are various sections of the course which were exposed to different stimulii, or different achievement level groups.

2.0 Item Selection Algorithm

As seen in the previous section, a teacher could use information about $P_{ab.}, a = 1,\cdots,A, b = 1,\cdots,B_a$ during each period. However, good estimates of $P_{ab.}$ require many test items, and there are only T units of testing time for each student during a period. Therefore, it is necessary to efficiently select the items to be used on the tests for each period. The algorithm to be developed will generate n_{ab}, the number of items to be used from MO_{ab}. The actual selection of items is then done on a stratified random basis, the stratification being done on the performance objectives within the MO. Thus, if $n_{ab} = 8$ and MO_{ab} contains 4 performance objectives, then 2 items will be selected at random from each performance objective within MO_{ab}.

Assume that there are S students in the course. Enough items can be selected to consume T units of time, and the same items given to all the students, or enough items can be selected to consume $S \cdot T$ units of time, and a different set of items given to each student. In general, one can select enough items to consume $L \cdot T$ units of time, $1 \leq L \leq S$, and give each set of items to S/L students. The following thoughts should be considered in the selection of L. When $L = 1$, then students can more easily be compared (all respond to the same items) than when $L = S$; while when $L = 1$, estimates of $P_{ab.}$ will contain more variance (due to sampling errors) than when $L = S$. If $L = S$, then a huge number of items is needed to consume $S \cdot T$ units of time, and the cost of producing the actual test forms can be enormous. Moreover, if each student takes a different set of items, then the item and student characteristics will be confounded. Section 4.0 examines the problem of obtaining simultaneous information about student and item characteristics. In present projects undertaken by the CAM staff with $S = 100$ to 150, L has ranged from

8 to 10. This value of L enables one to do considerable sampling within each MO, while still retaining much of the power for comparing groups of students.

What criteria can be used to select items from the various MO's? The present algorithm uses the following five.

a) The average time required to answer an item pertaining to MO_{ab}. Denote this by t_{ab}.

b) The value of the information (K_{ab}) on MO_{ab} during the current period. For instance, if MO_{ab} is not being taught for several months, then K_{ab} will be low, while if MO_{ab} was taught last week, then K_{ab} will be high. $\{K_{ab}, a = 1,\cdots,A, b = 1,\cdots,B_a\}$ is an input into the model and presently is viewed as being subjectively derived. Further theoretical developments may enable K_{ab} to be determined objectively.

c) The prior knowledge of $P_{ab.}$. If a teacher is quite confident that $P_{ab.}$ is very low, then there may be no need to use any items from MO_{ab}. This information is input into the model as a binomial prior (n',r'), where the prior expected value of $P_{ab.}$ is r'/n' and the variance is $(r'/n')(1-r'/n')/(n'+1)$. An excellent discussion of binomial priors is given in <u>Introduction to Statistical Decision Theory</u> (ISDT) [7].

d) The errors of sampling and measurement of items relating to MO_{ab}. See section 3.0.

e) A value function indicating the value of how close the estimate $\hat{P}_{ab.}$ is to the true value $P_{ab.}$. Denote this function by $V_{ab}(\hat{P}_{ab.}, P_{ab.})$. Note that $V_{ab}(\cdot,\cdot)$ depends upon the MO_{ab}. The value function to be used here will be of the squared-error loss type. This is not the only meaningful value function to use, but it appears to be the easiest from a computational viewpoint. The specific form of the function is (see figure 2),

$$V_{ab}(\hat{P}_{ab.},P_{ab.}) = -K_{ab}(\hat{P}_{ab.} - P_{ab.})^2$$

This value function is properly used when one is interested in an estimate of $P_{ab.}$ and does not have to choose one of a finite set of actions based upon the value $P_{ab.}$.

In ISDT it is shown for the case of pure binomial sampling (i.e., ignoring measurement errors) that the prior value of n_{ab} items from MO_{ab} is

$$I(n_{ab}) = K_{ab}\,\check{P}_{ab.}\,\frac{n_{ab}}{n_{ab} + n'_{ab}} \qquad n_{ab} = 0,1,2,\cdots$$

where

$$\check{P}_{ab.} = \frac{r'_{ab}}{n'_{ab}}\left(1 - \frac{r'_{ab}}{n'_{ab}}\right)/(n'_{ab} + 1)$$

is the variance of the prior estimate of $P_{ab.}$. Here the prior information is contained in the pair (n'_{ab},r'_{ab}). It can be shown that $I(n_{ab})$ is strictly concave in n_{ab} (see figure 3). The next step is to take into account the errors of measurement inherent to items measuring MO_{ab}. It is shown in section 3.0 that the relationship of the number of items from one MO that gives equivalent measurement errors with a fixed number of items from another MO is linear. Fixing one MO as a standard, one can generate the ratio of items (γ_{ab}) necessary to obtain equivalent information. This yields $J_{ab}(n_{ab}) \equiv I(\gamma_{ab}\, n_{ab})$ as the relative value of n_{ab} items from MO_{ab}. Finally, the problem to be solved to yield the desired values n_{ab} is

$$V_{ab}(\hat{P}_{ab.}, P_{ab.}) = -K_{ab}(\hat{P}_{ab.} - P_{ab.})^2$$

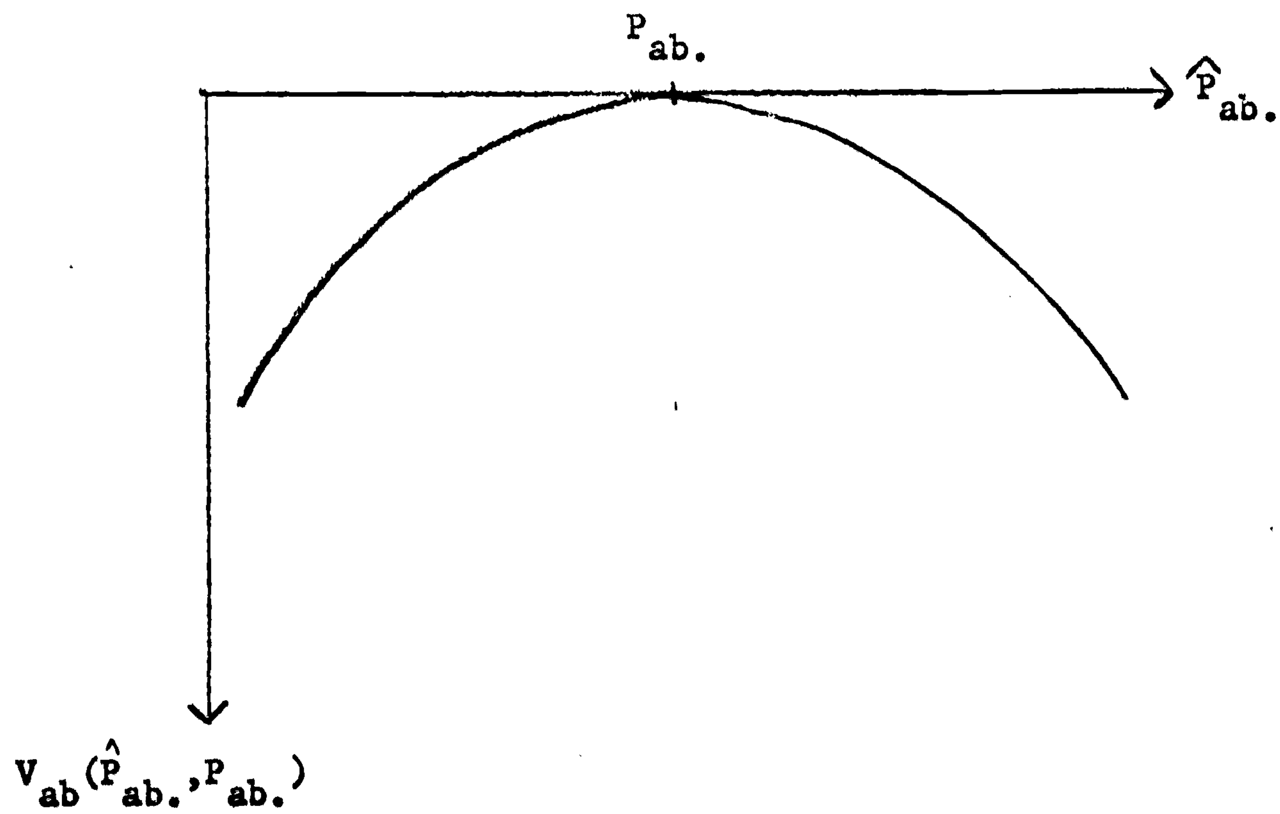

Figure 2

$$I(n_{ab}) = K_{ab}\,\check{P}_{ab.}\,\frac{n_{ab}}{n_{ab} + n'_{ab}}$$

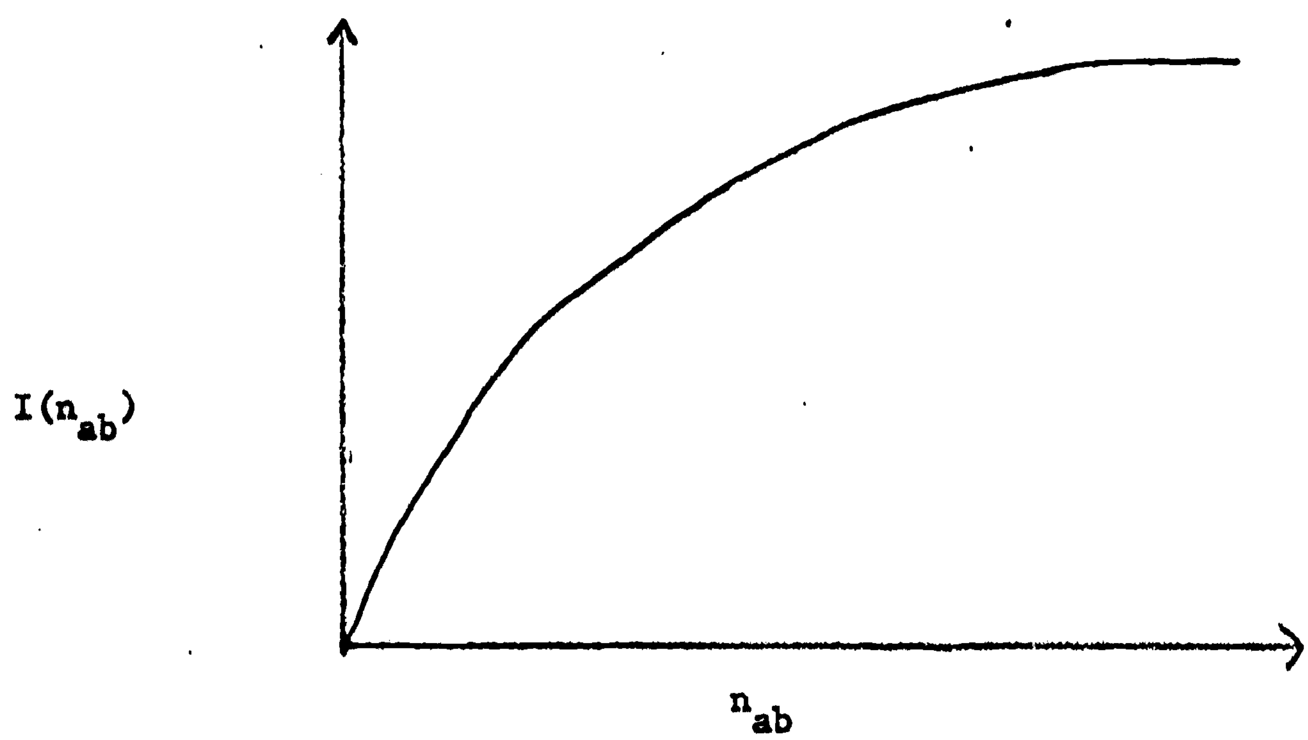

Figure 3

$$\max \sum_{a=1}^{A} \sum_{b=1}^{B_a} J_{ab}(n_{ab})$$

such that

$$\sum_{a=1}^{A} \sum_{b=1}^{B_a} t_{ab}\, n_{ab} \leq L \cdot T$$

$$n_{ab} \quad \text{integral}$$

Since $J_{ab}(n_{ab})$ is strictly concave in n_{ab}, this problem may easily be solved on the computer even for a very large number of MO's.

3.0 Errors of Measurement

In attempting to measure $P_{ab.}$, there will be two sources of errors that can be of different magnitude in different MO's. These are the format in which the items are written (free-response, multiple choice, or true-false) and the generic reliability of the test items. The following discussion will demonstrate the linear relationship between the number of items in each of two distinct MO's necessary to generate estimates with equivalent errors of measurement.

The question of the error variance introduced by the use of multiple choice versus free-response items is not an easy one to answer. There is the question as to whether a free-response format of a multiple choice item is the same question. Furthermore, what is the true guessing factor in a multiple choice item? If there are m alternatives to an item and the item is "perfectly" written, then if the student does not know the answer, he will respond correctly with probability $1/m$. However, how many teacher-constructed items are perfectly written in the sense described above? For the present analysis, assume that Sam and Joe are identical students (in the sense that they both know exactly the same material) and will respond to an n item test. However, Joe will answer in free-response format while Sam will be able to guess with probability $1/m$ if he does not know the correct response. The items given are drawn at random without replacement from the pool of all items that measure achievement in a given MO. Let X and Y be the number of correct responses given by Joe and Sam respectively, P the true percentage of the total item pool that Sam and Joe know, P_j the estimate of P from Joe's score and P_s the estimate of P from Sam's score. Then

$$\hat{P}_j = \frac{x}{n}$$

$$\sigma^2(\hat{P}_j) = \frac{1}{n} P(1-P)$$

Now $$\hat{P}_s = \frac{\frac{y}{n} - \frac{1}{m}}{1 - \frac{1}{m}}$$

as may be seen be writing $E(\frac{y}{n}) = P + (1-P)\frac{1}{m}$

and substituting y/n for $E(y/n)$ and $\hat{P}$ for P and

$$\sigma^2(\hat{P}_s) = \sigma^2\left(\frac{\frac{y}{n} - \frac{1}{m}}{1 - \frac{1}{m}}\right) = \frac{1}{(1 - \frac{1}{m})^2} \sigma^2(\frac{y}{n})$$

Since Sam's probability of a correct response for an item drawn at random is $P' = P + (1-P)\frac{1}{1/m}$,

$$\sigma^2(\frac{y}{n}) = \frac{1}{n}[P + (1-P)\frac{1}{m}] \cdot [1 - P - (1-P)\frac{1}{m}] \Rightarrow \sigma^2(\hat{P}_s) = \frac{1}{n} P(1-P) + \frac{1}{n}\frac{\frac{1}{m}(1-P)}{1 - \frac{1}{m}}$$

Thus, $\sigma^2(\hat{P}_s) - \sigma^2(\hat{P}_j) = \frac{1}{n}\frac{\frac{1}{m}}{1 - \frac{1}{m}}(1-P)$

Now one can ask how many test items does Sam have to respond to (n_s) in order to reduce $\sigma^2(\hat{P}_s)$ to $\sigma^2(\hat{P}_j)$ when Joe responds to n_j items.

$$\sigma^2(\hat{P}_j) = \frac{1}{n_j} P(1-P) = \frac{1}{n_s} P(1-P) + \frac{\frac{1}{m}}{n_s} \cdot \frac{1 - P}{1 - \frac{1}{m}} = \sigma^2(\hat{P}_s)$$

or $$n_s = n_j[1 + \frac{\frac{1}{m}}{(1 - \frac{1}{m})P}]$$

The above is the linear relationship necessary to correct for format differences. Note that this depends not only upon the number of alternative responses (m), but also on the true achievement level (P) of the students. The number of additional items necessary to produce sampling errors equivalent to free-response format when multiple choice format with m alternatives is used is $\frac{\frac{1}{m}}{(1-\frac{1}{m})P} n$. Note that as P approaches zero, this goes to infinity. So multiple choice and true-false items are especially inefficient for estimation purposes when used as pre-instruction measuring devices.

The above analysis yields a quite different conclusion from that of Lord [5]. In his work, he assumed that Sam and Joe both had the same expected score on their tests, i.e., $E(X) = E(Y)$, and came to the correct conclusion that the standard errors of their scores were equal. However, under his assumption, Sam and Joe would be different students with different knowledge levels, and therefore not comparable in a fasion Lord indicated.

There is no reason to believe that the generic reliability (or the generic error of measurement) will be the same from one MO to another. A discussion of these errors of measurement and methods of estimating them are presented in Lord and Novick [6], (Chapters 8 and 9), Cronbach [2], and Rajaratnam [8]. Once the generic reliabilities have been estimated, the Spearman-Brown prophecy formula can be used to derive the linear relationship necessary in the item selection algorithm to correct for measurement error between MO's. If MO_s and MO_r have generic reliabilities ρ_s and ρ_r respectively for the same number of items n, then the number of items (n_s) from MO_s needed to make ρ_s equal to ρ_r is given by

$$n_s = n \frac{\rho_r(1-\rho_s)}{\rho_s(1-\rho_r)} \quad .$$

4.0 Administration of Tests to Students

Once the $n_{ab}, a = 1,\cdots,A, b = 1,\cdots,B_a$ items have been selected from MO_{ab} as described in Section 2.0, the L "approximately" parallel forms must be constructed. This can be done by arranging the items on a one dimensional scale according to the chronological presentation to the class of the performance objective to which the items are related, and assigning every L^{th} item to the same form. Rigorously speaking, this will result in a violation of the time constraint for some of the forms. However, the variance among students' solution time for each item will be quite large, and unless most of the time consuming items are assigned to one form, no gross violations should result. Moreover, a few appropriate switches of items should correct any of these gross violations. Once the L forms have been constructed, each student in the course must be assigned a form to take. This assignment task should be done rigorously as shown below in order to get the most possible information out of the test data.

Consider two partitions of the students in the course. A partition is a collection of subsets of all the students such that each student is in one and only one subset. The two partitions that probably will be of interest are ones based upon achievement levels and physical attributes (different sections, teachers, etc.) of the class. For notational purposes, let $U_i, i = 1,\cdots,I$ and $V_j, j = 1,\cdots,J$ be the two partitions. Next, consider the cartesian product of the two partitions as shown below. Let $W_{ij} = U_i \cap V_j$ be those students in both U_i and V_j while $m_{ij} \geq 0$ is the number of students in subset W_{ij}. Finally, let $a_{ij\ell}, \ell = 1,\cdots,L$ be the number of students in W_{ij} who will respond to form ℓ. It is the $a_{ij\ell}$ that must be calculated. The requirements that all L forms are distributed evenly among partition V are

$$(4.1) \qquad \left| \sum_{i=1}^{I} a_{ij\ell} - \sum_{i=1}^{I} a_{ij\ell'} \right| \leq 1 \ \forall \ell,\ell' = 1,\cdots,L;\ j = 1,\cdots,J$$

while that of partition U are

$$(4.2) \qquad \left| \sum_{j=1}^{J} a_{ij\ell} - \sum_{j=1}^{J} a_{ij\ell'} \right| \leq 1 \ \forall \ell,\ell' = 1,\cdots,L,\ i = 1,\cdots,I$$

Once the $a_{ij\ell}$ are found that satisfy the above constraints, then one randomly selects a_{ij1} students from W_{ij} and gives them form 1, randomly selects another a_{ij2} and gives them form 2, etc. Present research is directed toward developing a computer algorithm to find a solution $a_{ij\ell}, i = 1, \cdots, I, j = 1, \cdots, J, \ell = 1, \cdots, L$.

The above procedure of assigning forms to students is useful for two main reasons. Firstly, if the forms are viewed as fixed and it is desired to estimate characteristics of the forms within the population of the class, a doubly stratified sample of students of approximate size S/L has been drawn at random from the class for each form. Thus, for a fixed item on a form, an unbiased estimate of its properties in the whole class is obtained, and this estimate has lower variance than a simple random sampling estimate due to the double stratification. The items are no longer confounded with any student or groups of students as defined by the partitions U and V. On the other hand, if one views the student groups as fixed and wishes to estimate the groups' performance on a given MO, a stratified (by performance objective ---Section 2.0) sample of items has been drawn within the MO. Furthermore, essentially the same items have been given to all groups within the same partition which enables one to make stronger statistical statements about differences between the groups. If different groups receive different sets of items, then an additional variance component is introduced into the comparison data, the variance being due to item differences. The actual statistical problems are slightly more complex than represented here due to the fact that within each subgroup (U_i or V_j) different students respond to different sets of items.

References

[1] Bloom, Benjamin S. (Ed.), Taxonomy of Educational Objectives, The Classification of Educational Goals, Handbook I, The Cognitive Domain, McKay Company, New York (1956).

[2] Cronbach, L. J., Rajaratnam, N., and Gleser, G. C., Theory of Generalizability: A Liberalization of Reliability Theory, British Journal of Statistical Psychology (1963), V. XVI, p. 137.

[3] Cronbach, L. J., and Gleser, G. C., Psychological Tests and Personnel Decisions, University of Illinois Press, Urbana (1965).

[4] Lord, F. M., Reliability of Multiple Choice Tests as a Function of Number of Choices per Item, Journal of Educational Psychology (1944), 35, 175-180.

[5] Lord, F. M., Do Tests of the Same Length Have the Same Standard Error of Measurement?, Educational and Psychological Measurement (1957), 17, 510-521.

[6] Lord, F. M., and Novick, M. R., Statistical Theories of Mental Test Scores, Addison-Wesley (1968).

[7] Pratt, J. W., Raiffa, H., Schlaifer, R., Introduction to Statistical Decision Theory, McGraw-Hill (1965).

[8] Rajaratnam, N., Cronbach, L. J., and Gleser, G. C., Generalizability of Stratified-Parallel Tests, Psychometrika (1965), 30, 39-56.

[9] Rurh, G. M., and Degraff, M. H., Corrections for Chance and "Guess" vs. "Do Not Guess" Instructions in Multiple-Response Tests, Journal of Educational Psychology (1926), Vol. XVII, 368-375.