THE USE OF LOG-LINEAR MODELS IN EVALUATING MASTERY PROGRAMS

Jennie P. Yeh

CSE Report No. 150
November, 1980

Center for the Study of Evaluation
Graduate School of Education, UCLA
Los Angeles, California 90024

The research reported herein was supported in whole or in part by a grant to the Center for the Study of Evaluation from the National Institute of Education, U.S. Department of Education. However, the opinions and findings expressed here do not necessarily reflect the position or policy of NIE and no official NIE endorsement should be inferred.

Table of Contents

INTRODUCTION

Conventionally, program effectiveness has been judged by comparing the average levels of achievement of two or more groups (e.g., hypothesis testing). The underlying assumption of this approach is that if a particular program is superior to another in its effectiveness, then students in the superior program should score higher than students who are not in the program. However, there are sound reasons why such an assumption is inappropriate for programs whose intended outcomes are either complete mastery or some minimal level of competency (i.e., results interpreted against a criterion). If two programs are compared and both fail to produce mastery or student achievement at the intended level, then both programs are ineffective, and it matters relatively little whether students in one program outperform those in another. Then, unlike program evaluation in which success is determined solely by level of achievement using, for instance, group means, it can be argued that the effectiveness of a mastery program should be judged by the number of students who obtain passing scores on a mastery measure. Given the above rationale, there exists a need for alternative statistical methods for evaluating program effects in terms of mastery. A number of statistical limitations also support this need:

In mastery testing, the metrics are not necessarily continuous. The

outcomes of a mastery test can be considered not as a "score," but rather as a "sign" (Harris, 1974). That is, all individuals classified as masters receive a sign, e.g., (+) and all individuals classified as non-masters receive a different sign, e.g., (-), no matter what score they receive on the test. Another problem lies in the distribution of mastery. While the structure of mental abilities has been studied extensively, the structure of proficiencies is relatively unknown. It has been established, however, that in measuring mastery the tester is not dealing with the natural organization of the mind (Cronbach, 1960). How difficult an item is for a given individual is strictly a function of his experience. Bloom (1968) has suggested that training affects the contours of distributions of true scores. It may be reasonable to assume that before training occurs, the aptitude to learn a skill is normally distributed and the amount of knowledge possessed by most individuals with respect to that specific skill is very little. Mastery learning theory asserts that under appropriate instructional conditions, virtually all students can master most of what they are taught (Block & Buros, 1976), and that student test scores will cluster around the higher end of the continuum. Therefore, the distribution of test scores may be negatively skewed.

When testing for significance of group differences, if the underlying distributions from the two populations are not of the same shape but are symmetrical, we encounter little difficulty. If they differ in skewness, however, the distribution of obtained t's also has a tendency to be skewed, with a greater percentage of obtained t's falling outside of one limit than the other. This tends to bias probability statements (Boneau,

1960). Glass, Peckham and Sanders (1972) also report that skewed populations can seriously affect the level of significance and power of directional, or one-tailed, tests.

Another problem arises when we examine the basic assumptions required for hypothesis testing. For instance, the assumption of homogeneity of variance required by ANOVA would be difficult to meet if various treatment groups differ in the amount of learning each has acquired. As was previously stated, before learning takes place, the subject matter of most students' knowledge clusters around the nonmastery point. After learning has taken place, most individuals will have mastered the skill, and their abilities will cluster around the high end of the continuum of scores. In either case, the variance is small. It is at the halfway point of the training program that individuals' scores tend to spread out (Boneau, 1966). Violating the assumption of homogeneity of variance has very slight effect on α (type I error), if the sample sizes are equal, although actual α always seems to be slightly increased over the nominal α. With unequal sample size, however, α may be seriously affected. As reported in Glass (1973), Scheffé found, for example, in a case where the nominal α is .05, $n_1 = 15$ and $n_2 = 5$ and $\sigma_1^2 = .2\sigma_2^2$, the actual probability of obtaining a significant t-ratio when the null hypothesis is true is .178. Thus, one is nearly three and one-half times more likely to commit a type I error than is supposed.

The effects of nonnormality and heterogeneous variance were studied by Norton (1953) and Boneau (1960). Both studies used only cases with equal sample size. The results, as summarized by Glass, et al., suggest that nonnormality and heterogeneous variance appear to combine additively to affect either level of significance or power.

This paper presents an alternative method for dealing with the above limitations. The proposed method assumes a mastery assessment model with the following characteristics described in terms of the construction and scoring of mastery tests as well as the organization and reporting of group testing results:

1) For program evaluation, the skill that is being tested must be explicitly defined in terms of a performance criterion. Tests designed to assess mastery of a single skill consist of a collection of test items which are highly homogeneous in content and form so that each item response provides an unbiased estimate of the examinee's mastery status with respect to that skill.

2) When multiple skills are tested, the knowledge possessed by an individual missing an item testing skill A is not comparable to that of an individual missing an item testing skill B. Thus, items constructed to test each skill should be considered as constituting a unique test. The result of these item groups should be reported in a manner such that the mastery standing on each skill can be clearly determined. Results from a multiskill measure should be reported in the form of a vector having as its elements the mastery decisions representing each skill tested.

3) Regardless of the number of outcome categories ($m \geq 2$) (e.g., master/nonmaster or other combinations), each testee can be assigned, on the basis of his/her score, to one of several mutually exclusive categories with respect to a single skill. On a test of multiple ($n \geq 2$) objectives, an individual will then be assigned to one of the m^n possible categories. For example, on a test of two skills, a testee can be assigned to one of two categories with respect to each skill ($m = 2$); therefore,

there are 4 (2^2) possible categories to which an individual can belong: He can master the first skill but fail the second skill; he can master the second skill but fail the first skill; he can master both skills; he can fail both skills.

The following is an example of how test results are reported according to the model:

Suppose a reading program on literal comprehension is designed to achieve the following objectives:

a. The learner will identify, in a reading selection, the explicitly-stated main idea.

b. The learner will write a summary of a passage he has just read.

c. Given a reading selection in which a question is posed and the answer explicitly stated, the learner will identify the answer to the question.

d. After reading a given selection, the learner will identify the correct sequence of its main events or concepts.

Test items reflecting these objectives are then chosen to assess the success or failure of the reading program. Ten items are randomly selected for assessing each objective. Mastery is defined as responding correctly to 8 out of 10 items or making 20%, or less, incorrect responses. Thus, if Student A correctly answers 7 items for objective a, 8 for b, 4 for c, and 10 for d, she will be assigned to the mastery group on objectives b and d. The numeral 0 is assigned to designate nonmastery and 1 for mastery. Student A's responses, therefore, may be summarized as a vector with entries (0,1,0,1). Student B answers 6, 6, 10, and 7 items correctly on objectives a to d, and his response pattern, using mastery scoring, will appear as

(0,0,1,0). By examining each individual's response vector, it is possible to ascertain which skills have, and have not, been mastered. This type of data provides specific information on each individual's mastery status for each objective and thereby yields a more complete diagnosis than does a single numeral score. It is assumed here, that mastery levels are not arbitrary and that two people with the same ability will attain the same mastery score.

The model can also be applied to group data. For example, to compare the effectiveness of two teaching methods (A and B) in achieving student mastery of the previously listed objectives, students are randomly assigned to two instructional groups. One group uses method A (Group A) and the other group uses method B (Group B). Results obtained from measurements that are designed to test mastery of these objectives can be arranged as shown in Table 1.

In Group A only 62 students master all four objectives, whereas in Group B, 122 students master the four objectives. Three hundred and five students in Group A and 217 in Group B master objective a but do not master objectives b, c, and d.

The data obtained from the use of this model requires a statistical approach different from the conventional approach used with continuous data. With continuous data, the independent variables in least-square analyses may be either categorical or continuous or both. When the response variables (dependent variables) are qualitative (categorical), such as the data presented in Table 1, however, the statistical analyses change fundamentally, because the random variable in the statistical model is discrete and must be described by a discrete probability distribution.

TABLE 1

Summary of Responses of
Two Method Groups on Four Objectives

Response Vectors Objectives				Number of persons with given mastery pattern	
a	b	c	d	Group A (Control)	Group B (Experimental)
1	1	1	1	62	122
1	1	1	0	70	68
1	1	0	1	31	33
1	1	0	0	41	25
1	0	1	1	283	329
1	0	1	0	253	247
1	0	0	1	200	172
1	0	0	0	305	217
0	1	1	1	14	20
0	1	1	0	11	10
0	1	0	1	11	11
0	1	0	0	14	9
0	0	1	1	31	56
0	0	1	0	46	55
0	0	0	1	37	64
0	0	0	0	82	53

Although in large samples these distributions can be approximated by the continuous univariate and multivariate normal distribution, the approximations are often inaccurate in practical-sized samples and cannot be recommended in general.

The following section discusses an alternative statistical approach--one which is appropriate for the analysis of the type of complex contingency table data presented in Table 1.

THE LOG-LINEAR MODEL

The study of complex contingency tables stems from the early work of Bartlett (1939) on contingency table interaction. His work was followed by that of Fisher (1943) and Norton (1945). More recently, this area of inquiry has been extended by such scholars as Plackett (1962), Good (1963), Goodman (1970, 1973), and Bock (1969, 1973).

Bock and Goodman were concerned with the development of general asymptotic procedures for testing higher order interactions in complex contingency tables. The methods used by these two researchers apply generally in the analysis of multifactor, multiresponse data when the response is qualitative but not ranked. The solutions offered by both yield similar results (Goodman, 1973). The procedure for establishing the model matrix, using Bock's method, is straightforward and can easily be generalized to various experimental designs. Bock's solution, obtained through his method of estimation, in many respects parallels the general linear model approach to the univariate and multivariate analysis of variance applied to continuous data (Graybill, 1961).

The following is a description of Bock's Log-Linear Model applied to the analysis of multivariate quantitative data. In this description, if

each testee in the sample is assigned to one of several mutually exclusive categories, as in the case of a single skill test, the category frequencies thus generated will be called single-response data. If each subject makes a number of responses, and each response is assigned to one of several mutually exclusive categories as in the case of multiple skills testing, the resulting frequencies will be called multiple response data.

The Two-Category (Single-Response) Model

The problem of characterizing the relationship between a structured independent variable and a qualitative response variable is relatively simple when there are two response categories, e.g., masters and nonmasters. Data in this form are referred to as dichotomous or binary.

Let us suppose that the experimenter establishes experimental conditions in which the responses of the subjects are observed. These conditions differ in physically identifiable or measurable ways which are hypothesized to influence the probability of response. Let N_j subjects be observed under experimental condition j. The subjects are then scored dichotomously according to the presence or absence of some response, in this instance "mastery" or "nonmastery." Let the number of subjects under experimental condition j (j = 1,2,...n) who show the response be r_{j1} and the number of subjects who fail to show the response be r_{j2}. Let the corresponding response proportions be P_{j1} and P_{j2}; then $P_{j1} + P_{j2} = 1$. The probability of observing the frequencies r_{j1} and r_{j2} among N_j randomly-selected subjects is given by the binomial law:

$$(1) \quad P(r_{j1},r_{j2}|N_j) = \frac{N_j!}{r_{j1}!\, r_{j2}!} P_{j1}^{r_{j1}} P_{j2}^{r_{j2}}$$

Considering the group j data separately, it can be shown that the best unbiased estimates of P_{j1} and P_{j2} are respectively:

(2) $P_{j1} = r_{j1} / N_j$;

$P_{j2} = r_{j2} / N_j$

But if there is a functional relationship between P_{j1} and an independent variable, $\underline{x}_j$, it may be possible to estimate the population proportions more accurately by estimating the parameters of this relationship rather than by using (2). The experimenter's problem is, then, to predict these proportions by estimating response probabilities P_{j1} and P_{j2}, expressed as a function of the independent variables which determine the experimental conditions.

To avoid computational difficulties arising from inadmissible estimates, and in the hope of simplifying the relationship between the response probabilities and the independent variables, logistic transformation of the probabilities is typically performed. In the binomial case, the logistic response law is defined by:

(3) $P_{j1} = \frac{e^{Z_j}}{1 + e^{Z_j}}$

or $\log (p_{j1} / P_{j2}) = Z_j$

The quantity Z_j is called a binomial logit. Where there is a single quantitative independent variable, a polynomial model may be suitable:

$$(4)\quad Z_j = \beta_o + \beta_1 x_j + \beta_2 x_j^2 + \ldots + \beta_{r-1} x_j^{r-1}$$

or, in matrix notation:

$$(5)\quad \underset{nx1}{Z} = \underset{nxr}{X}\ \underset{rx1}{\underline{\beta}}$$

Model (5) parallels the univariate regression model. In that vector Z corresponds to the expected values of the observations, the matrix X contains the known value of the independent variable, and $\underline{\beta}$ is the vector of coefficients to be estimated.

The Multiple Response Model

In the multiple response situation, let the number of categories in the response classification m be equal to or greater than 2 ($m \geq 2$). Let the number of subjects under experimental condition j ($j = 1,2,\ldots n$) who fall in category h of the classification be r_{jh}, where $\sum_j^m r_{jh} = N_j$. Assume random sampling of subjects, so that the probability of the response frequencies r_{jh} is given by the product multinomial:

$$(6)\quad \prod_j^n \frac{N_j!}{r_{j1}!\, r_{j2}! \ldots r_{jm}!} P_{j2}^{r_{j2}} \ldots P_{jm}^{r_{jm}}$$

To express the response probabilities as functions of the experimental variables, let us introduce the multivariate logits of group j as $\underline{z}_j = (z_{j1}, z_{j2}, \ldots, z_{jm})$ and generalize the logistic response law as follows:

$$(7)\quad P_{j1} = e^{z_{j1}}/D_j \;;$$

$$P_{j2} = e^{z_{j2}}/D_j \;;$$

$$\vdots$$

$$P_{jm} = e^{z_{jm}}/D_j \;;$$

$$\text{where } D_j = e^{z_{j1}} + e^{z_{j2}} + \ldots + e^{z_{jm}}$$

In establishing a linear model connecting the logits with the independent variables, we must provide for the possibility that structure of the categories is implied in the response classification. A model for the logits sufficiently general to include both a structured response classification and multiple experimental variables is as follows:

$$(8)\quad \underset{nxm}{Z} = \underset{nxq}{X} \cdot \underset{qxt}{\beta} \cdot \underset{txm}{A}$$

where Z is the logistic transformation of the response proportions. The rows of Z are $\underline{z}_j$ $(j = 1,2,\ldots n)$.

The matrix X is a matrix of the known values of q independent variables associated with each of the subject groups. In the present model, it accounts for variation in the response probabilities over the experimental conditions. Given the data presented in Table 1, if the response probabilities of the two instructional groups differ, the basis for this part of the model (called the physical part) should be rank 2. For example:

$$X = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

where column 1 accounts for the grand mean and column 2 accounts for the difference between the two experimental groups.

If the groups do not differ, a rank 1 model comprising only the first column of this matrix will suffice.

$\underline{\beta}$ contains unknown parameters of the effect, and the role of matrix A in the model is to account for variation of the response probabilities across categories (m) of the response classification. The categories of the response classification in the present example are represented by the different response vectors shown in Table 1. Thus there are 16 categories in the "response" part of the model. In general, for k dichotomized responses, there are 2^k categories in the response part of the model.

The rank of the response part of the model depends on the number of categories and ways of classification in the contingency table, and on how the response probabilities are assumed to be determined. In this instance, the alternatives for establishing the response matrix are: a rank 4 main category model, a rank 10 first-order interaction model, or a rank 4 second-order interaction model. The basis consists of as many rows from the conventional matrix of single degree of freedom contrasts for a 2^k factorial design excluding the vector corresponding to the grand mean. The grand mean vector is excluded because the response probabilities are invariant under change of location of the logits. A main category model is given:

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \end{bmatrix}$$

In the analysis of the data from Table 1, a full rank model would consist of a rank 2 physical part and rank 15 response part. Using the full rank model, the estimated response probabilities would be equal to the observed probabilities, and no gain in precision would result from fitting the model.

If the subject groups have a crossed or nested structure, X will be the model matrix for the design and will, in general, be of deficient rank. Suppose the rank of X is $r \leq q$. Then it will be necessary to reparameterize the model by setting:

$$X = \underset{nxr}{K} \cdot \underset{rxq}{L} ,$$

the rows of L are coefficients of linearly estimatable functions of the parameters in the columns of β. That is, L must be of rank r and must satisfy the usual condition for linear estimability:

$$\text{rank} \begin{bmatrix} X \\ \cdots \\ L \end{bmatrix} = \text{rank } [X] = \text{rank } [L] = r$$

Like X, matrix A will also frequently be of less than full rank. Suppose rank $A = s \leq t$, then it will also be necessary to reparameterize the model for factoring A as:

$$A = \underset{txs}{S} \cdot \underset{sxm}{T}$$

where rank (S) = s and the columns of S are linearly dependent on those of A.

After these reparameterizations, (8) becomes:

(9) $$Z = K(LBS)T$$

$$= \underset{nxr}{K} \cdot \underset{rxs}{\Gamma} \cdot \underset{sxm}{T}$$

The matrix K is the "pre-factor," Γ, the parameter, and T the "post-factor," after Roy's (1957) terminology.

Using the multinomial probability function (6), estimates of the model (9), parameters can be obtained, given that the model is identified, through the use of the method of maximum likelihood. This method gives asymptotically best estimates in that they are asymptotically normally distributed, consistent, and, in general, have asymptotically more efficient standard errors than other classes of estimators (Rao, 1965).

Let us define, in matrix form, the vectors of response frequencies and response probabilities for experimental condition j to be:

$$\underset{\sim}{r}_j = \begin{bmatrix} r_{j1} \\ r_{j2} \\ \cdot \\ \cdot \\ \cdot \\ r_{jm} \end{bmatrix} \quad \text{and} \quad P_j = \begin{bmatrix} P_{j1} \\ P_{j2} \\ \cdot \\ \cdot \\ \cdot \\ P_{jm} \end{bmatrix}$$

Assuming random assignment of samples of subjects to the n groups, the likelihood equations may be expressed as:

$$(10)\quad \underset{rsx1}{\underline{G}(\Gamma)} = \sum_{j}^{n} \underset{sxm}{T}(\underset{rx1}{\underline{r}_j} - \underset{mx1}{N_j\underline{P}_j}) \times \underset{rx1}{\underline{K}_j} = \underline{0} \ ,$$

where x denotes the Kronecker product and $\underline{K}_j$ is the j-th row of K written as a column. Successive elements in this expression represent derivatives taken with respect to r_{kh}, with the second subscript varying first.

Let us define the mxm matrix:

$$\begin{bmatrix} P_{j1}(1 - P_{j1}) & - P_{j1}P_{j2} & \cdots & - P_{j1}P_{jm} \\ -P_{j2}P_{j1} & P_{j2}(1 - P_{j2}) & \cdots & - P_{j2}P_{jm} \\ \vdots & \vdots & & \vdots \\ -P_{jm}P_{j1} & -P_{jm}P_{j2} & & P_{jm}(1 - P_{jm}) \end{bmatrix}$$

The rs x rs matrix of second derivatives may then be expressed as:

$$(11)\quad -H(\Gamma) = - \sum_{j}^{n} N_j T W_j T' \times \underline{K}_j \underline{K}_j'$$

To obtain the maximum likelihood estimates, the Newton-Raphson iterative procedure is carried out as follows: Starting at initial estimates for Γ, called $\hat{\Gamma}_t$, such that:

$$(12)\quad \hat{\Gamma}_{t+1} = \hat{\Gamma}_t + \delta_t \ ,$$

where the t subscript denotes the iterative step, and

$$\delta_t = H^{-1}(\hat{\Gamma}_t) G(\hat{\Gamma}_t) \ ,$$

where $H^{-1}(\hat{\Gamma}_t)$ is the negative of the inverse of the matrix of second derivatives evaluated at $\hat{\Gamma}_t$. The process may be repeated until the corrections (δ) vanish.

The theory of maximum likelihood estimation establishes that such estimates are consistent, i.e., they converge to population values as the sample size becomes indefinitely large; and their joint distribution is approximated by the multivariate normal distribution with mean equal to the population value, and variance-covariance matrix equal to the negative inverse of the matrix of second derivatives of the likelihood function:

$$(13)\quad V(\hat{\Gamma}) = H^{-1}(\Gamma)$$

Large sample standard errors for the estimated parameters of the logistic model are obtained by extracting the diagonal of the matrix of second derivatives in the final iteration of the Newton-Raphson solution.

Tests of Goodness-of-Fit

The goodness-of-fit of the model to the data, at the final values of the parameters, can be tested by the chi-square approximation for the likelihood ratio statistics:

$$(14)\quad \chi^2_{L.R.} = 2\sum_{jk}^{nm}\sum r_{jk}\ln(r_{jk}/N_j\hat{P}_{jk})$$

The $\hat{P}$'s are the expected response probabilities computed from the maximum likelihood estimates of the parameters in the hypothesized model. The number of degrees of freedom for this chi-square is the difference between the number of parameters fitted when the observed proportions estimate directly the population proportions and the number fitted in the model:

d.f. = $n(m-1) - rs$.

In moderate size samples, the Pearsonian chi-square,

$$(15)\quad \chi^2_P = \sum_{jk}^{nm} \frac{(r_{jk} - N\hat{P}_{jk})^2}{NP_{jk}}$$

and the likelihood ratio chi-square usually differ only slightly. However, Fisher suggests the L. R. chi-square is more appropriate when some of the expected values are small.

In the next section, the fictitious data presented in Table 1 will be analyzed using the model, as will a set of data taken from an actual evaluation study.

APPLICATION OF THE MODEL

Example 1

The set of data which comprises Table 1 was originally used by Solomon and has been analyzed by both Goodman (1973) and Bock (1969). Use of this data allows for a more complete investigation of the important attributes of the proposed method of analysis than would be possible using empirical data alone.

In the previous discussion, a hypothetical reading program was described. Its objectives fell into four areas:

a. Identifying main ideas

b. Writing summaries

c. Selecting correct answers to questions about material read

d. Identifying correct sequences of events

Let us suppose the teachers reported that, in their judgment, the program was effective in teaching students to master the selected skills, that it was

decided that a test should be constructed to assess student mastery on these skills, that such a test was constructed and given. We can see that student responses to such a test could be used to further improve the test. For example, to study how well the instrument is able to differentiate masters and nonmasters, the test could be administered to a random sample of students who went through the program and a random sample of students who did not.

Results from this hypothetical test are presented in Table 1. Response vector (1 1 1 1) indicates the number of students who mastered each of the four objectives: 62 students in the control group mastered these objectives, and 122 of the experimental students also achieved mastery. Response vector (0 0 0 0) indicates the number of students who were classified as "nonmasters" on any of the four objectives.

Several alternatives must be considered in setting up the model for the analysis. If the response probabilities of the two groups differ, the basis for the physical part of the model should be rank 2. For example:

$$X = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

There are three alternatives for the response part of the model. These alternatives are a rank 4 main-category model, a rank 10 first-order interaction model, or a rank 14 second-order interaction model. The possible bases for the response part of the model are given by the appropriate single degree of freedom contrasts for 2^4 factorial design.

For the data that is presented in Table 1, the goodness-of-fit chi-square for several alternative models are shown in Table 2.

TABLE 2

Goodness-of-Fit Tests
for Various Models

Rank of physical part	Rank of response part	L.R. chi-square	Degrees of freedom	Probability
1	4	188.3615	26	<.0005
2	4	130.4839	22	<.0005
1	10	74.4613	20	<.0005
2	10	11.5628	10	.40<p<.30

The rank 2, rank 10 model appears to fit the data well (χ^2 = 11.5628). The estimated parameters using this model are shown in Table 3. The Group A + Group B effect reflects the general response probability of each subtest for the corresponding issue. Subtest a has high response probability (β=.75) and subtest b (β=-.76) has low probability, and thus the response patterns beginning with 1-0 (mastery of objective a, nonmastery of objective b) are especially frequent. These estimated parameters can be used in much the same way as item difficulties of classical test theory. In general, many students mastered (objective a) main idea and a few mastered (objective b) writing summaries. Patterns of responses with mastery of objective c also tend to be more probable, and this tendency is increased by the interactive effect of a joint occurrence of mastery of objective c with mastery of objective a or b.

TABLE 3

Estimated Parameters and Standard Errors
for Rank 2, Rank 10 Model

Response Effect	Physical effect A + B Estimated Parameter	A + B Standard Error	A - B Estimated Parameter	A - B Standard Error
a	.7542	(.03133)	.0433	(.06266)
b	-.7636	(.03195)	.0233	(.06390)
c	.1473	(.02984)	-.1949	(.05967)
d	-.0120	(.02913)	-.2825	(.05827)
a x b	-.0316	(.03116)	-.1033	(.06232)
a x c	.1240	(.02468)	-.0061	(.04936)
a x d	.0289	(.02476)	.0907	(.04952)
b x c	.1515	(.12550)	-.0427	(.05101)
b x d	.0500	(.02427)	-.0839	(.04853)
c x d	.1053	(.01894)	.0222	(.03790)

The estimated contrasts between Group A and Group B indicate that the difference between groups is almost entirely due to differences in their response probabilities or individual subtests (specifically c and d). The very small values of the interaction contrasts as compared to their standard errors suggest little difference in pairwise association between items from one group to the other. Notice that the two groups are most differentiated on issues c and d. Both the experimental and control groups have mastered objective a (β = .7542 corresponds to an estimated probability of correct response of 0.82). There is no difference between the level of mastery of the experimental and control groups on objective a (β = .0433 which is not statistically different from zero). This finding also illustrates the fact that measurements which have medium response probability are better discriminators than measurements like subtests a and b, which have extreme (low or high) response probability.

Table 4 gives the estimated cell frequencies of the rank 2, rank 10 model, which can also be used to check the goodness-of-fit of the model.

TABLE 4

Response of Students in Member and Nonmember Groups

Objective responses a b c d	Observed frequencies		Expected frequencies	
	Control	Experimental	Control	Experimental
1111	62	122	68.6	120.9
1110	70	68	62.6	66.5
1101	31	33	29.7	33.7
1100	41	25	43.2	27.0
1011	283	329	276.3	322.4
1010	253	247	260.6	256.2
1001	200	172	201.4	179.1
1000	305	217	302.7	207.4
0111	14	20	11.6	21.5
0110	11	10	14.3	11.0
0101	11	11	8.2	10.0
0100	14	9	16.0	7.5
0011	31	56	33.5	62.2
0010	46	55	42.6	46.2
0001	37	64	39.7	57.4
0000	82	53	80.2	62.2
Total	1491	1491	1491.2	1491.2

Example 2

The following example illustrates the use of mastery measures in evaluating educational programs with pre- and post-testing. This example is based on the 10th-grade results of the Los Angeles Unified School District's ESEA Title III Ecology Program.

The purpose of the program was to develop a package of learning activities that would effectively teach students to master selected ecologically related facts and concepts. Students from the 10th-grade biology classes in the district were randomly selected to participate in the program. One half of these students were randomly assigned to the experimental group and were given instructions based on the newly developed 10th-grade Ecology Learning Activity Module, while the remaining half were assigned to the control group and continued in their regular programs. To ascertain the relative merits of the new module compared to the regular program, a test consisting of several subtests was constructed. Each subtest contained items measuring a specific objective and was administered to both groups of students prior to, and immediately following, the ten-week instructional period.

In the program, mastery was defined as being able to answer at least 75% of the items on a subtest correctly.

The frequency distribution of attainment of mastery on two subtests is presented in Table 5. The concepts (designated as M and N) measured by these two subtests are:

M: Man, living in different cultural settings throughout history, has viewed his position in nature in a variety of ways. These range from living his role as a fragile functional unit, to his believing he has the divine right to develop, to possess, and to destroy.

N: Modern man's values about how the land should be used have led to large-scale exploitation and ruin of wild areas. Some of the ways man destroys land are by clear-cutting, paving, strip-mining, and dam-building.

TABLE 5

Frequency Distribution of Attainment of Mastery on Pre- and Post-testing on Subtests M and N

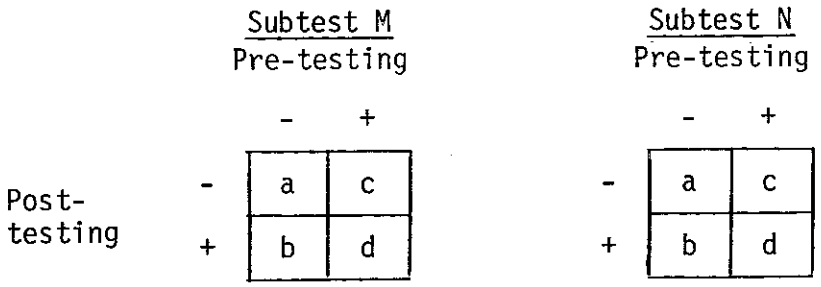
Subtest M		Subtest N			
Pre	Post	Pre	Post	Experimental	Control
0	0	0	0	28	19
0	0	0	1	9	8
0	0	1	0	8	10
0	0	1	1	7	12
0	1	0	0	10	9
0	1	0	1	19	5
0	1	1	0	7	9
0	1	1	1	24	11
1	0	0	0	4	11
1	0	0	1	3	10
1	0	1	0	8	5
1	0	1	1	8	16
1	1	0	0	20	24
1	1	0	1	28	18
1	1	1	0	14	17
1	1	1	1	93	48

With two experimental conditions, the physical part of the model is set to be:

$$X = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

where column 1 accounts for the grand mean and column 2 accounts for the difference between the two experimental groups.

Each individual's mastery status with respect to responses M and N can be classified into one of the following four categories:



a. Non-mastery on both pre- and post-testing.

b. Non-mastery on pre-testing and mastery on post-testing.

c. Mastery on pre-testing but nonmastery on post-testing.

d. Mastery on both testing occasions.

With respect to the two skills, therefore, there are 16 (4 x 4) possible categories of response classifications. These categories are represented by the different response vectors in Table 5. The response matrix resembles a conventional 4^2 factorial design excluding the vector corresponding to the grand mean. A full rank model (rank 15) consists of three degrees of freedom (4-1=3) for each for the two main categories, and 9 (3 x 3 = 9) degrees of freedom for the first-order interaction. The vector for each single degree of freedom should be constructed according to the hypotheses the experimentor wishes to test. For instance, suppose that the following vectors are chosen to form the basis of the response:

$$A = \begin{array}{c} \begin{array}{cccccccccccccccc} & & & & & & \multicolumn{9}{c}{\text{Column}} & & & & \\ & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & \\ \end{array} \\ \left[\begin{array}{cccccccccccccccc} 1 & 0 & -1 & 1 & 0 & -1 & 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 1 & \\ 1 & 0 & -1 & -1 & -1 & 0 & -1 & -1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & \\ 1 & 0 & -1 & -1 & 1 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & \\ 1 & 0 & -1 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & -1 & \\ & & & & & & & & & & & & & & & \\ -1 & 1 & 0 & 1 & 0 & -1 & -1 & 0 & 1 & 1 & 0 & -1 & 0 & 0 & 0 & \\ -1 & 1 & 0 & -1 & -1 & 0 & 1 & 1 & 0 & -1 & -1 & 0 & 0 & 0 & 0 & \\ -1 & 1 & 0 & -1 & 1 & 0 & 1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & \\ -1 & 1 & 0 & 1 & 0 & 1 & -1 & 0 & -1 & 1 & 0 & 1 & 0 & 0 & 0 & \\ & & & & & & & & & & & & & & & \\ -1 & -1 & 0 & 1 & 0 & -1 & -1 & 0 & 1 & -1 & 0 & 1 & 0 & 0 & 0 & \\ -1 & -1 & 0 & -1 & -1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & \\ -1 & -1 & 0 & -1 & 1 & 0 & 1 & -1 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & \\ -1 & -1 & 0 & 1 & 0 & 1 & -1 & 0 & -1 & -1 & 0 & -1 & 0 & 0 & 0 & \\ & & & & & & & & & & & & & & & \\ 1 & 0 & 1 & 1 & 0 & -1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & -1 & \\ 1 & 0 & 1 & -1 & -1 & 0 & -1 & -1 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & \\ 1 & 0 & 1 & -1 & 1 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & \\ 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 1 & \end{array}\right] \end{array} \begin{array}{c} \\ \text{Row} \\ 1 \\ 2 \\ 3 \\ 4 \\ \\ 5 \\ 6 \\ 7 \\ 8 \\ \\ 9 \\ 10 \\ 11 \\ 12 \\ \\ 13 \\ 14 \\ 15 \\ 16 \end{array}$$

The rows of the matrix correspond to the 16 categories of the response classification, and each column corresponds to 1 degree of freedom. The first 3 columns account for the response classifications on subtest M alone. Columns 4 through 6 account for the response classifications on subtest N. Columns 7 through 15 are the interaction terms which, in the logistic model, correspond to association between the interacting classification; estimates of these effects serve as measures of the direction and extent of association.

If a full rank model--rank 2 physical part and rank 15 response part--is fitted, the estimated response probabilities would be equal to the observed probabilities. Residual chi-squares can be computed by fitting a rank 1 response part rank 1 physical part model using column 1 of the physical matrix (X) and column 1 of the response matrix. By adding columns 2,3,4,...,14 successively to the basis of the response matrix, residual chi-squares can be obtained for models disregarding the difference

in response patterns between the two experimental conditions. The difference due to sampling (experimental condition) can be taken into account by adding column 2 of the sample matrix (X) to the model. Table 6 presents the residual chi-squares for models of various ranks. The difference between the respective residuals then provides chi-squares corresponding to each component.

In the response matrix, column 1 tests the probability of change in mastery status between pre- and post-testing versus the probability of no change on M. Column 4 tests the same hypothesis with respect to subtest N. Columns 2 and 5 test the probability of positive change versus negative change on M and N, respectively. Columns 3 and 6 test the probability of persistent nonmasters versus students who knew the skills from the beginning.

Since column 1 of the sample matrix is the grand mean, all models with 1 degree of freedom in the sample matrix give the general tendency of the response pattern. All models with 2 degrees of freedom in the sample matrix give the interaction of treatment group membership and response pattern.

General Tendency

In Table 6, the first 6 response components, with the exception of component 5, are significant at the $\alpha = .05$ level, and thus the following conclusions can be drawn:

1. With respect to skill M, disregarding the difference in instructional group membership,
   a. The component χ^2 of 209.03 indicates that there is a greater tendency for students to remain in the same mastery status

TABLE 6

Test of Goodness-of-Fit
Ecology Program Tenth Grade Results

Ranks Sample	Response	Likelihood ratio residual chi-square	df	Component chi-square	df
1	0	464.90	30		
1	1	255.87	29	209.03*	1
1	2	250.55	28	5.32*	1
1	3	176.60	27	73.95**	1
1	4	122.88	26	53.72**	1
1	5	120.15	25	2.73	1
1	6	94.14	24	26.01**	1
1	7	88.64	23	5.50*	1
1	8	88.64	22	0.00	1
1	9	88.64	21	0.00	1
1	10	88.53	20	0.11	1
1	11	87.89	19	0.64	1
1	12	87.79	18	0.10	1
1	13	86.85	17	0.94	1
1	14	85.64	16	1.21	1
1	15	65.37	15	10.37**	1
2	0	430.38	30		
2	1	254.82	28	175.56**	2
2	2	236.90	26	17.92**	2
2	3	161.20	24	75.70**	2
2	4	107.20	22	54.00**	2
2	5	102.11	20	5.09	2
2	6	72.42	18	29.69**	2
2	7	65.29	16	7.13*	2
2	8	65.13	14	0.16	2
2	9	65.08	12	0.05	2
2	10	64.69	10	0.39	2
2	11	55.08	8	9.61**	2
2	12	54.93	6	0.15	2
2	13	54.10	4	0.83	2
2	14	52.91	2	1.19	2
2	15	Null	0	52.91**	2

**significant at $\alpha = .01$.

*significant at $\alpha = .05$.

rather than to change (.70 vs. .30, see Table 7). The pre-test marginals indicate that 63% (Table 7) of the students had already mastered the material prior to instruction. A program that results in a much higher probability may be wasteful, while one that results in a lower probability may need strengthening.

TABLE 7

Contingency Table of Pre-test by Post-test
(Skill M)

		Pre-testing: Nonmaster	Pre-testing: Master	
Post-testing	Non-master	101 (.19)	65 (.12)	166 (.31)
	Master	94 (.18)	262 (.51)	356 (.69)
		195 (.37)	327 (.63)	

b. The tendency of positive change is greater than negative change ($\chi^2 = 5.32$). Assuming that an individual cannot unlearn a skill between pre- and post-testing, the negative change can be used as an index for measurement error. The significant χ^2 thus enables us to conclude that while it is more likely that students will remain in the same mastery status rather than change, the probability of a positive change is nevertheless greater than error.

The likelihood of mastery on both testing occasions is greater than for nonmastery on both occasions. Taking points a and b

into account, it can be concluded that the material is too easy for students at the 10th grade level.

2. With respect to skill N, disregarding the difference in instructional group membership,

   a. The component χ^2 value of 73.95 indicates that there is a greater tendency for students to remain in the same mastery status rather than to change (.66 vs. .36, see Table 8). This tendency again is due to the heavy concentration of students in the master master cell.

TABLE 8

Contingency Table of Pre-test by Post-test
(Skill N)

		Pre-testing: Nonmaster	Pre-testing: Master	
Post-testing	Non-Master	125 (.24)	78 (.15)	203 (.39)
	Master	100 (.19)	219 (.42)	319 (.61)
		225 (.43)	297 (.57)	

   b. Although there is a greater tendency for students to change from nonmaster to master, the χ^2 value of 2.73 indicates that it is not statistically significantly different from error.

   c. There are more consistent masters than persistent nonmaster (χ^2 = 26.01).

3. Component 7 is the interaction term of components 1 and 4. (In the response matrix [p. 26], column 7 is the product of columns 1 and 4.) This component indicates an association between subtests M and N; those who are likely to change their mastery status on M are also likely to change their status on N, while those who do not change on M will not change on N. The χ^2 value of 5.50 for this component is significant at $\alpha = .05$ level. In the following table (Table 9) we see that 60% of the cases concentrated in the two cells on the main diagonal. Further, there is a disproportionate concentration of cases in the lower right-hand cell, which is mainly caused by the large percentage of students who mastered the two skills on both pre- and post-testing.

TABLE 9

Contingency Table of Change by Subtest
(Subtest M)

		Change	No change	
Subtest N	Change	66 (.13)	112 (.21)	178 (.34)
	No change	93 (.18)	251 (.48)	344 (.66)
		159 (.31)	363 (.69)	

Differentiating between
Experimental and Control Groups

When fitting a rank 2 sample matrix with a response matrix of various ranks, the component chi-squares test the hypothesis of how much response component differentiates the two experimental groups. Examination of the lower half of Table 6 shows that of the 15 components, 8 are statistically significant. On the basis of these tests, the following conclusions are drawn:

1. With respect to skill M:

   a. There is a greater tendency for experimental students to change positively than for control students (.21 vs. .15, see Table 10). The chi-square for this component is 175.15.

TABLE 10

Contingency Table of Treatment by
Pre-test by Post-Test
(Subtest M)

Pre-test		Experimental		Control	
		Nonmaster	Master	Nonmaster	Master
Post-test	Nonmaster	52 (.18)	23 (.08)	49 (.21)	42 (.18)
	Master	60 (.21)	155 (.53)	34 (.15)	107 (.46)
		112 (.39)	178 (.61)	83 (.36)	149 (.64)

This hypothesis is of primary interest in this example. An effective educational program should be one that can make a meaningful change in students. The traditional method of

post-test scores, or a significant difference between the experimental and the control group, does not provide sufficient information for making such an educational decision as whether a curriculum that can increase the reading score by 1 point on the test should be adopted when such an increase has been shown to be significant.

Or, to provide another example, let us assume that two programs (A and B) teach skill T. On the average, students in program A answer 30% of the items correctly on the test measuring mastery of skill T, while for those in program B the average is 15%. The scoring criterion is such that to be considered a master of the skill, one must mark at least 70% of the items correctly. Even if students in program A score significantly higher than students in program B, neither teaches many students to master skill T and, therefore, we may not wish to adopt either program.

b. The component chi-square value of 17.92 indicates that the probability for students to remain in the same status is greater for the experimental group than for the control group. To investigate the source of this effect, let us examine Table 10. At pre-testing, the two groups are comparable in that 61% of the experimental and 64% of the control students are classified as masters. However, 53% of the experimental students are masters on both occasions, while only 46% of the control students are masters on both occasions. The error due to measurement (defined before as master on pre-testing but nonmaster on post-testing) is greater

for the control group (.18) than for the experimental group (.08), suggesting a greater tendency for control students to guess on post-testing.

c. The difference in percentage between consistent masters and persistent nonmasters is greater in the experimental group than in the control group. The chi-square value for this component is 75.70. Examination of Table 10 shows that in the experimental group of 112 nonmasters, 60 became masters (54%) on post-testing; in the control group, 41% of the nonmasters on pre-testing became masters on post-testing. As discussed in (b) above, of those who were classified as masters on pre-testing as a result of guessing, there is a greater percentage of the experimental group who became masters after instruction.

2. With respect to skill N:

a. The probability for students to remain in the same status is greater for the experimental students than for the control group ($\chi^2 = 54.00$). This probability is primarily a result of the disproportionate concentration of students in the consistent masters cell for the experimental group (.46). Notably, of those who are classified as masters at pre-testing, 78% of the experimental students are classified as masters upon post-testing. In the control group, only 67% of the masters on pre-testing are classified as masters on post-testing (Table 11).

TABLE 11

Contingency Table of Treatment by
Pre-test by Post-test
(Subtest N)

Post-test \ Pre-test	Experimental		Control	
	Nonmaster	Master	Nonmaster	Master
Nonmaster	62 (.21)	37 (.13)	63 (.27)	41 (.18)
Master	59 (.20)	132 (.46)	41 (.18)	87 (.37)
	121 (.41)	169 (.59)	104 (.45)	128 (.55)

b. Of those who change their mastery status, the experimental group shows greater probability for learning, while for the control group the probability of change due to learning is equal to measurement error. The χ^2 (5.09) statistic for this hypothesis falls a little below α = .05 (.057).

c. As on subtest M, the difference in percentage between consistent masters and persistent nonmasters on subtest N is also greater in the experimental group than in the control group (χ^2 = 9.69).

3. Interaction terms:

a. Component 7 of the response matrix is the interaction term of components 1 and 4. The significant χ^2 (7.13) indicates that the association between subtest M and subtest N is different between the two student groups. In the experimental group, 64% of the students are found in cells a and d (Table 12); and only 57% of students in the control group appear in these cells. It is more

likely for the experimental students who change mastery status on M also to change on N, and for those who are unchanged on M to remain unchanged on N. While control students may learn about M or N on their own, they may learn about one but not the other and, therefore, their mastery status may change on one skill but not the other. Experimental students who are exposed to the instruction of M are also exposed to the instruction of N, which therefore results in the higher association.

TABLE 12

Contingency Table of Change by Subtests by Treatment

Subtest M		Experimental		Control	
		Change	No change	Change	No change
Subtest N	Change	a 37 (.13)	c 59 (.20)	a 29 (.13)	c 59 (.23)
	No change	b 46 (.16)	d 148 (.51)	b 47 (.20)	d 103 (.44)
		83 (.29)	207 (.71)	76 (.33)	162 (.67)

* * * * *

While the models presented above provided the opportunity to test a number of hypotheses, none is found to fit the data well. An alternative approach to the analysis of the Ecology Program data would be to treat the response classification as four main categories: skill M pre-test, skill M post-test, skill N pre-test, and skill N post-test. There are two

mutually exclusive classes, masters and nonmasters, within each main category. The response matrix, thus, would appear the same as the design matrix of 2^4 factorial design without the vector corresponding to the grand mean:

$$X = \begin{array}{c} \begin{array}{cccccccccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & \end{array} \\ \left[\begin{array}{cccccccccccccccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & -1 & -1 & -1 & 2 \\ 1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & -1 & -1 & 3 \\ 1 & 1 & -1 & -1 & 1 & -1 & -1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 & 4 \\ 1 & -1 & 1 & 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & -1 & -1 & 5 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 & 6 \\ 1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & 7 \\ 1 & -1 & -1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & 8 \\ -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & -1 & 9 \\ -1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 & -1 & 1 & 10 \\ -1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & 11 \\ -1 & 1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & 1 & -1 & 1 & -1 & 12 \\ -1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & 1 & 13 \\ -1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & 1 & -1 & 14 \\ -1 & -1 & -1 & 1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 & 15 \\ -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 16 \end{array}\right] \\ \begin{array}{cccccccccccccccc} A & B & C & D & AB & AC & AD & BC & BD & CD & ABC & ABD & ACD & BCD & ABCD & \end{array} \end{array}$$

Each of the first four columns tests the probability of falling into the mastery versus nonmastery class on each of the four main categories. Columns 5 through 10 test the two-way interaction of the four main categories. The last six columns are the three-way and four-way interactions. Table 13 presents the goodness-of-fit for models of various ranks.

TABLE 13

Goodness-of-Fit for Models
of Various Ranks

Ranks		Likelihood		
Sample	Response	chi-square	df	Probabilities
1	10	44.48	20	0.0013
2	4	179.56	22	0.0000
2	10	6.09	10	0.8075

The rank 2 rank 10 model has a very good fit. Using this model, the estimated β's and the standard errors for β's are presented in Table Again, the first column of the sample matrix indicates the general tendency of the two groups combined, while the second column compa. two groups. The findings are listed below.

TABLE 14

Estimated Effects and Standard Errors
for Rank 2, Rank 10 Model
(Ecology Program)

Response effect (β)	Physical effects	
	Experimental + Control	Experimental - Control
A	-0.255 (.105)*	0.452 (.211)
B	-0.593 (.106)*	-0.648 (.212)*
C	-0.081 (.103)	0.078 (.201)
D	-0.174 (.105)	-0.216 (.210)
AB	1.426 (.216)*	0.270 (.432)
AC	0.439 (.206)	1.013 (.412)
AD	0.471 (.214)	-0.527 (.428)
BC	0.332 (.219)	-0.333 (.438)
BD	0.752 (.219)	1.570 (.437)*
CD	1.146 (.199)	-0.054 (.399)

*2.5 times greater than its standard error as suggested by Goodman (1973).

General Tendency

1. β_A β_B indicate that there is a greater percentage of masters than nonmasters with respect to M on both pre- and post-testing.
2. Students' mastery status tends to remain the same rather than to change on both M (β_{AB} = 1.426) and N (β_{CD} = 0.752) and those who are nonmasters on one skill tend to be nonmasters on the other skill.

Comparing the Two Experimental Groups

1. Upon post-testing, there are more masters in the experimental group than in the control group on subtest M (β_B = -0.648). As shown in Table 15, 74% of the experimental students have mastered M on post-testing, while only 61% of the control students have mastered M for post-testing (Table 15).

TABLE 15

Frequency Distribution and Percentage of Masters and Nonmasters on Four Testings

	Experimental		Control		Total	
	Master	Non-master	Master	Non-master	Master	Non-master
Subtest M Pre-testing	178 (.61)	112 (.39)	149 (.64)	83 (.36)	327 (.62)	195 (.38)
Subtest M Post-testing	215 (.74)	75 (.26)	141 (.61)	91 (.39)	356 (.68)	166 (.32)
Subtest N Pre-testing	169 (.58)	121 (.42)	128 (.55)	104 (.45)	297 (.57)	225 (.43)
Subtest N Post-testing	191 (.66)	99 (.34)	128 (.56)	104 (.44)	319 (.61)	203 (.49)

2. Upon pre-testing, the experimental students' mastery on M and N is related. This association, however, does not occur in the control group. Since the students are randomly selected and then randomly assigned to the two instructional groups, it is not possible to determine the cause of this difference on the pre-testing (Table 16).

TABLE 16

Contingency Table of Pre-testing Results on Subtests M and N by Treatment Groups

Subtest M	Experimental		Control	
	Nonmaster	Master	Nonmaster	Master
Subtest N: Nonmaster	66 (.23)	55 (.19)	41 (.18)	63 (.27)
Subtest N: Master	46 (.16)	123 (.42)	42 (.18)	86 (.37)
	112 (.39)	178 (.61)	83 (.36)	149 (.64)

3. χ^2_{BD} indicate that there is a stronger positive relation between the mastery status on M and N for the experimental group than for the control group. In the experimental group, it is more likely for those who have mastered M also to master N, and for those who have not mastered M not to master N. This finding may be a result of the fact that the experimental students are exposed to instruction on both M and N, whereas the control students may be exposed to one but not the other. In other words, the experimental students' learning experience on the two skills is planned, whereas the control students' exposure to these two skills is not (Table 17).

TABLE 17

Contingency Table of Mastery Status on M and N by Treatment Groups

Subtest M		Experimental		Control	
		Nonmaster	Master	Nonmaster	Master
Subtest N	Nonmaster	48 (.17)	51 (.18)	45 (.19)	59 (.25)
	Master	27 (.09)	164 (.57)	46 (.20)	82 (.35)
		75 (.26)	215 (.75)	91 (.39)	141 (.60)

Table 18 gives the observed and estimated frequencies using the rank 2 rank 10 model. It can also be used to check the fit of the model.

TABLE 18

Observed and Estimated Response of Students in Experimental and Control Groups

M_1	M_2	N_1	N_2	Observed Frequencies Experimental	Observed Frequencies Control	Estimated Frequencies Experimental	Estimated Frequencies Control
0	0	0	0	28	19	26	19
0	0	0	1	9	8	8	8
0	0	1	0	8	10	9	9
0	0	1	1	7	12	9	13
0	1	0	0	10	9	13	10
0	1	0	1	19	5	19	4
0	1	1	0	7	9	5	9
0	1	1	1	24	11	23	11
1	0	0	0	4	11	7	11
1	0	0	1	3	10	3	10
1	0	1	0	8	5	6	5
1	0	1	1	8	16	7	15
1	1	0	0	20	24	16	23
1	1	0	1	28	18	29	19
1	1	1	0	14	17	17	18
1	1	1	1	93	48	93	48
			Total	290	232	290	232

DISCUSSION

This study presents a proficiency assessment model that provides information on individuals' mastery status on specific skills. The assessment model describes the construction and scoring of mastery tests as well as the organization and reporting of individual and group testing results.

When the model is used, each individual's mastery status (mastery or nonmastery) with respect to several skills can be represented by a vector with n elements, where n equals the number of skills tested. This type of multivariate quantitative data is analyzed by the log-linear model which incorporates both a sampling factor (e.g., experimental group versus control group) and a response structure. The model, therefore, provides test statistics to assess the differences in capability of the sampled groups as well as the structure represented by the response factors.

The two examples presented illustrate the applicability of the log-linear model in the analysis of proficiency assessment results. In the first example, the contingency table from Solomon's (1961) study was adapted for use in a hypothetical setting with the intention of illustrating how to evaluate the differentiating capability of master measures.

A good instrument, theoretically, should be one that can accurately differentiate between "experts" and "nonexperts" and between students who have had instruction and students who have had no instruction. To evaluate the "goodness" of the Reading Comprehension Test empirically, the hypothetical example presumes that the test was given to a random sample of program students and to a random sample of nonprogram students.

The results of the statistical analysis indicates that the instrument as a whole differentiates between the experimental and control groups. The two subtests measuring objectives c and d differentiate between the two groups the most while the subtest for objective a does not differentiate between the two groups. Measures with medium response probability, overall, are better discriminators than measures with extreme response probability. However, with mastery measures, the groups of experts could all have a score of 1, and the group of nonexperts could all have a score of 0. When the number of cases in the two groups is equal, therefore, such as in the given example, the overall response probability would be .50. In this context, the above-mentioned conclusion agrees with the classical theory of measurement.

The log-linear model, however, is capable of detecting differences in the mastery status of instructed and uninstructed groups when there is a disproportion of students in the groups which would lead to overall response propensities outside the middle range. Thus, if there are four times as many experts as nonexperts, and all the experts demonstrate mastery while the nonexperts are nonmasters, then the overall response probability will be .80; but the log-linear model will still detect the group difference.

The second example deals with the use of mastery measures in an actual educational program evaluation. Here, the 10th grade results of the Los Angeles Unified School District's ESEA Title III Ecology Program were analyzed, using the proposed model. Two different approaches were employed in formulating the response matrix.

The results of the analysis indicated that students using the program material were more likely to improve than students not using the material, on the two skills taught in the program. Among the group of students with instruction, those who mastered one skill were more likely to master the other skill; this relationship did not exist among the group of students without instruction. This result could be interpreted to mean that the students in the noninstructed group could be exposed to one skill but not the other, whereas students in the instructional group were exposed to both skills. Therefore, in the noninstructed group, students' mastery status on either skill was related to the degree of advantageous exposure they received; while in the instruction group, exposure to both units seemed to have a reinforcing effect.

The results also indicated that students' mastery status in both groups was more likely to remain the same than to change. This finding indicated that the program was not efficient, primarily as a result of the great percentage of students who were masters on both skills at pre-testing (.62 and .57). As suggested by Novick and Lewis (1974), a carefully monitored program will, typically, be such as to suggest a prior probability distribution that assisgns a probability of just more than .50 to the region above the criterion level. By this rule, the post-testing goal was achieved upon pre-testing in the given example. According to Novick and Lewis, then, the program may be wasteful and should have been revised for implementation at a lower grade level as soon as the pre-testing results were reported.

The log-linear model presented in this study permits a concise analysis of data arising from the use of mastery measurements. Further explorations of the use of this model should concentrate on manipulations of the mastery level criteria in light of the efficiency notion proposed by Novick and Lewis, as applied in educational evaluations.

REFERENCES

Bartlett, M. S. Contingency table interactions. Journal of the Royal Statistical Society, 1935, 2, 248-252.

Block, J. H., & Burns, R. B. Mastery learning. In L. Shulman (Ed.), Review of research in education, Vol 4. Itasca, IL: Peacock, 1976.

Bock, R. D. Estimating multinomial response relations. In R. C. Bose, et al. (Eds.), Contributions to statistics and probability: Essays in memory of S. N. Roy. Chapel Hill: University of North Carolina Press, 1969.

Bock, R. D. Multivariate statistical methods in behavioral research. New York: McGraw-Hill, 1974.

Bock, R. D., & Jones, L. V. The measurement and prediction of judgment and choice. San Francisco: Holden Day, 1966.

Bock, R. D., & Yates, G. Multiqual log-linear analysis of nominal or ordinal qualitative data by the method of maximum likelihood. Chicago: National Educational Resources, 1973.

Bloom, B. S. Test reliability for what? Journal of Educational Psychology, 1942, 33, 517-526.

Bloom, B. S. Learning for mastery. UCLA Center for the Study of Evaluation. Evaluation Comment, 1968, 1(2), 1-12.

Boneau, C. A. The effects of violations of assumptions underlying the t test. Psychological Bulletin, 1960, 57, 49-64.

Carroll, J. B. The nature of the data, or how to choose a correlation coefficient. Psychometrika, 1961, 26(4).

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

Conover, J. Practical nonparametric statistics. New York: John Wiley, 1971.

Coulsen, J. E., & Cogswell, J. F. Effects of individualized instruction on testing. Journal of Educational Measurement, 1965, 2, 59-64.

Cronbach, L. J. Essentials of psychological testing, 3rd ed. New York: Harper and Row, 1970.

Fisher, R. A., & Roberts, J. A. F. A sex difference in blood-group frequencies. Nature, 151, 640-641.

Glaser, R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 514-541.

Good, I. J. Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. Annual Mathematical Statistics, 1963, 34, 911-934.

Goodman, A. Guided and unguided methods for the detection of models for a set of T multidimensional contingency tables. Journal of the American Statistical Association, 1973, 68, 165-175.

Goodman, L. A. The analysis of multidimensional contingency tables: Stepwise procedures and direct estimation methods for models for multiple classifications. Techometrics, 1970, 12.

Graybill, A. An introduction to linear statistical models. New York: McGraw-Hill, 1961.

Harris, C. W. Some technical characteristics of mastery tests. CSE Monograph Series in Evaluation, 1974, 3, 98-115.

Keesling, J. W. Empirical validations of criterion-referenced measures. CSE Monograph Series in Evaluation, 1974, 3, 159-176.

Norton, J. W. Calculation of chi square from complex contingency tables. Journal of the American Statistical Association, 40, 251-259.

Plackett, R. L. A note on interactions in contingency tables. Journal of the Royal Statistical Society, B24, 162-166.

Rao, C. R. Linear statistical inference and its applications. New York: John Wiley, 1965.

Rao, S. N. Some aspects of multivariate analysis. New York: John Wiley, 1957.