Analysis of Test Results via Log-Linear Models

Frank B. Baker and Michael J. Subkoviak
University of Wisconsin

The recently developed log-linear model procedures are applied to three types of data arising in a measurement context. First, because of the historical intersection of survey methods and test norming, the log-linear model approach should have direct utility in the analysis of norm-referenced test results. Several different schemes for analyzing the homogeneity of test score distributions are presented that provide a finer analysis of such data than was previously available. Second, the analysis of a contingency table resulting from the cross-classification of students on the basis of criterion-referenced test results and instructionally related variables is presented. Third, the intersection of log-linear models and item parameter estimation procedures under latent trait theory are shown. The illustrative examples in each of these areas suggest that log-linear models can be a versatile and useful data analysis technique in a measurement context.

A new data analysis technique known as log-linear models has been developed over the past decade, providing a means for the analysis of qualitative data at a level of sophistication that has long been available for quantitative data. Under the log-linear procedures, a researcher can establish a linear model for the observed frequencies in the cells of a multidimensional contingency table in a manner similar to that used in the analysis of variance. The intent is to find a log-linear model that (1) fits the data and (2) is parsimonious in the number of terms employed. A variety of levels of presentation of this approach are provided in recent books by Bishop, Fienberg, and Holland (1975), Bock (1975, chap. 8), Everitt (1977), Fienberg (1977), and Haberman (1978, 1979). In addition, a number of computer programs such as MULTIQUAL (Bock & Yates, 1973), ECTA (Fay & Goodman, 1975), and 3F of the BMD/P series (Dixon & Brown, 1979) implement the technique.

The log-linear models methodology arose primarily within the context of survey research where the interest was in understanding the interrelationships among qualitative variables used to define a multidimensional contingency table. It has long been recognized that the norming of educational and psychological measuring instruments rests heavily upon the sampling plans and data analysis techniques of survey research. For example, both Lord (1959) and Angoff (1971) have discussed the survey sampling aspects of the test norming process. However, the data analysis aspects of the norming process has been limited mainly to estimating population means and their standard errors.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 5, No. 4, Fall 1981, pp. 503–515
© Copyright 1981 Applied Psychological Measurement Inc.
0146-6216/81/040503-13$1.65

The purpose of this paper is to provide the reader with an introduction to the use of log-linear models within the context of educational and psychological measurement rather than to provide a comprehensive coverage of the full range of applications. Because of the historical intersection of survey research and educational measurements, the new capabilities represented by log-linear models have direct applications in the analysis of norm-referenced test results. In addition to this application, there are two other areas within psychometrics where the log-linear model approach can be employed. This paper will present an application based upon the results of criterion-referenced testing that provides an interesting analysis of such data. Finally, the work of Bock (1970, 1972, 1975) provides a direct link between log-linear models and latent trait theory: Consequently, two examples of estimating item parameters via log-linear models are provided.

Background

To introduce the logic underlying log-linear models, the classical analysis of a two-dimensional contingency table will be examined. If a sample of interest were categorized on the basis of two factors, say race (A) and item response (B), each with two levels, then the corresponding contingency table is that shown in Table 1. Upon dividing each cell entry by *N*, the proportion, P_{jk}, of the sample falling in each cell is obtained. The null hypothesis, of no relationship between race and item response (independence of the row and column factors), is given by

Table 1
Population of Examinees Cross-Classified on the Basis of Race and Item Response

	Item		Marginal
Group	Correct	Incorrect	Total
Black	f_{11}	f_{12}	$f_{1\cdot}$
White	f_{21}	f_{22}	$f_{2\cdot}$
Total	$f_{\cdot 1}$	$f_{\cdot 2}$	N

$$H_0: \quad p_{jk} = p_{j\cdot}\ p_{\cdot k} \qquad [1]$$

$$\text{for all } j = 1, 2, \ldots r$$

$$\text{and } k = 1, 2, \ldots c.$$

Since the parameters, p_{jk}, are unknown, their estimates can be obtained from the marginal totals,

$$\hat{P}_j = f_{j\cdot}/N, \quad \hat{P}_{\cdot k} = f_{\cdot k}/N \quad \text{and} \quad \hat{P}_{jk} = (\hat{P}_{j\cdot})(\hat{P}_{\cdot k}) \ . \qquad [2]$$

Then, under an assumption of independence,

$$E_{jk} = N(\hat{P}_{j\cdot})(\hat{P}_{\cdot k}) \qquad [3]$$

is an estimator of the expected frequency $F_{jk} = Np_{jk}$. A procedure for testing the null hypothesis can be established using the usual

$$\chi^2 = \sum_{j}^{r} \sum_{k}^{c} \frac{(f_{jk} - E_{jk})^2}{E_{jk}} \quad \text{with } (r - 1)(c - 1) \text{ degrees of freedom,} \qquad [4]$$

where r and c denote the number of rows and columns, respectively. From the above, it is clear that the classical analysis is based upon a multiplicative model, since under the null $F_{jk} = Np_{j\cdot}\, p_{\cdot k}$. Basically, what the log-linear model approach does is to convert the classical analysis from a multiplicative to a linear model. This is done by expressing the natural logarithms of the expected frequencies F_{jk} in terms of a linear model (see Everitt, 1977, for the associated algebra). In the case of a two-dimensional contingency table and the hypothesis of independent row and column classificatory factors, the log-linear model would be

$$\log F_{jk} = U + U_{1(j)} + U_{2(k)} \qquad [5]$$

Using Anova terminology, this model has the following meaning imparted to the U terms:

U is a constant common to all cells;

$U_{1(j)}$ is the contribution of the j^{th} level of the first main effect (the row factor here); and

$U_{2(k)}$ is the contribution of the k^{th} level of the second main effect (the column factor here).

Although this model closely parallels those used in Anova, even to the allocation of degrees of freedom, there are two important differences. First, the U terms are expressed in a logarithmic metric. Second, log-linear models do not employ an explicit error term. In fact, adding a term that represents the association of the row and column factors would yield a model exactly reproducing the cell frequencies. Such deterministic models are called "fully saturated" models in the present context. Adding this term to Equation 5 would result in

$$\log F_{jk} = U + U_{1(j)} + U_{2(k)} + U_{12(jk)} \qquad [6]$$

where $U_{12(jk)}$ is the contribution of the association of the j^{th} level of Factor 1 with the k^{th} level of Factor 2 to the logarithm of the cell frequency.

This general log-linear model representation can be extended to multidimensional contingency tables that can involve main effects as well as first, second, and higher order associations. For example, a log-linear model for the cell frequency in a three-factor contingency table could be

$$\log F_{ijk} = U + U_{1(i)} + U_{2(j)} + U_{3(k)} + U_{12(ij)} + U_{13(ik)}$$
$$+ U_{23(jk)} + U_{123(ijk)} \qquad [7]$$

where the fifth through seventh terms are the first-order associations of the classificatory variables and the last term is the second-order association. An important characteristic of such models is that they are hierarchical, that is, if a higher order association (such as $U_{123(ijk)}$) appears in a model, the corresponding lower-order associations and main effects also must be present. This assumption enables using a shorthand notation to represent a particular model rather than writing out the full expression. For example, the notation [123] represents Equation 7 above. The model

$$\log F_{ijk} = U + U_{1(i)} + U_{3(k)} + U_{13(ik)} \qquad [8]$$

would be represented by [13].

The log-linear model approach fits successive models of various numbers of terms to the observed frequency data in an attempt to find the most parsimonious model that accounts for the data. The goodness of fit between the observed cell frequencies, f_{jk}, and the expected frequencies, E_{jk}, yielded by the data under a given model, i.e., the "residual," is usually indexed by the following approximation to the likelihood ratio statistic:

$$G^2 = 2 \sum_{j}^{r} \sum_{k}^{c} [f_{jk} \ln (\frac{f_{jk}}{E_{jk}})] \quad \text{which is distributed as chi-square} \qquad [9]$$

with the degrees of freedom being rc less the number of terms fitted in the model. The difference between the residual G^2 statistic for two successive models, termed the component G^2, indexes the contribution of the additional terms in the more complex model to the goodness of fit.

Some Illustrative Examples

Analysis of Test Norm Data

Example 1. One form of analysis employed in test norming is to compare the test performance of subgroups of interest. Historically, this has been done via tests of equality of group means and/or variances as well as goodness-of-fit tests between pairs of group distributions. The log-linear model approach, however, enables the simultaneous testing of the homogeneity of entire test score distributions for multiple groups. For example, let it be assumed that the norm population can be stratified into three subgroups on the basis of race (black, white, and other). Suppose further that the test score has been partitioned arbitrarily into seven mutually exclusive and exhaustive score intervals. The result will be a 3 × 7 contingency table, such as presented in Table 2, in which the cell frequencies would be the number of examinees of a given racial group falling into each of the test score intervals. The hypothesis of interest is whether the score distributions are independent of the race factor. The appropriate log-linear model for testing this hypothesis as denoted by [1], [2] is given by

$$\log F_{jk} = U + U_{1(j)} + U_{2(k)} \ . \qquad [10]$$

This model was fit to the observed frequency data of Table 2 via the MULTIQUAL program, and there was not a good fit to the data ($G^2 = 32.5$ with 12 df, $p = .001$). Thus, the factors of race and test

Table 2
Test Score Distribution of Examinees
As a Function of Race

Race	Test Score							
	1	2	3	4	5	6	7	
Observed Frequency								
Black	7	59	243	380	243	63	5	1000
White	9	91	363	574	363	91	9	1500
Other	16	66	233	290	162	31	2	800
Observed Proportion								
Black	.007	.059	.243	.380	.243	.063	.005	1.000
White	.006	.061	.242	.383	.242	.061	.006	1.000
Other	.020	.082	.291	.363	.202	.039	.003	1.000

score were not independent. Inspection of Table 2 shows that the score distributions of the black and white groups are quite similar, while differing from the other group distribution. This suggests the interaction may be a function of this difference.

Example 2. Let it be further assumed that the demographic data of Example 1 also includes a second stratification factor of school location (urban, rural). As shown in Table 3, the stratification scheme now consists of a fully crossed 3 × 2 factorial layout with race as Factor 1 and location as Factor 2. The response layout remains the single factor, test score, with 7 levels and denoted as Factor 3. In this example, it is of interest to determine if the test score factor is independent of race, school location, and of the interaction of the race and school location factors.

Table 3
Number of Examinees At Each Test Score Level As
A Function of Race and School Location

Race and Location[a]	Test Score Level 1	2	3	4	5	6	7	n
Black								
U	4	25	106	195	124	36	3	493
R	3	34	137	185	119	27	2	505
White								
U	3	40	172	297	187	50	7	756
R	6	51	191	277	176	41	2	744
Other								
U	6	23	110	155	76	12	1	383
R	10	43	123	135	86	19	1	417

[a] U denotes urban location and R denotes rural location.

These concerns can be formulated via a series of log-linear models within a single analysis of the data using 3F of the BMD/P package, which is summarized in Table 4. The fully saturated model would be [123] and the next model of interest would be [12] [13] [23], which involves all possible first-order associations among the factors. The component G^2 yielded by comparing these two models would indicate whether a second-order association exists among race, location, and test score. The obtained value was $G^2 = 7.95$ with 12 *df* and $p = .789$, indicating no significant second-order association.

Table 4
Goodness of Fit Results By Successive Models For
Data of Table 3

Model[b]	Residual G^2	Residual df	Residual p	Component G^2	Component df	Component p
[123]	0	0	1.0			
[12][13][23]	7.95	12	.789	7.95	12	.789
[13][23]	8.71	14	.849	.76	2	.999
[1][2][3]	59.16	32	.002	50.45	18	.000

[b] Factor 1 is race, Factor 2 is location, Factor 3 is test score.

The next model of interest is [13] [23], which omits the association of race and location. The model yielded a residual $G^2 = 8.71$ with 14 *df* and $p = .85$; thus, the model fit the data. The component G^2 yielded by the difference between models [12] [13] [23] and [13] [23] was .76, with 2 *df* and $p = .999$, showing that the association between race and location has little impact upon the fit of the model. The main-effects-only model, [1] [2] [3], yielded a residual $G^2 = 59.16$ with 32 *df*, $p = .002$, indicating that this model does not fit the data. Comparison of this model with the model [12] [13] [23] yielded a component $G^2 = 51.21$ with 20 *df*, $p = .000$, and it can be concluded that first-order associations among the three factors exist. Comparing the main-effects-only model and [13] [23] produced a component $G^2 = 50.45$, with 18 *df* and $p = .000$.

From these results, three conclusions can be drawn. First, there is no second-order association among race, school location, and test score. Second, a main-effects-only model does not fit the data. Third, a model containing the association of race with test score, school location with test score, but not the association of race and school location, is appropriate.

Results such as those presented in Table 4 involve two considerations that are worth noting at this point. First, when a large number of models are fit to a set of data, the α level of the significance tests is an issue. Bock (1975, chap. 8) recommends dividing the α level by the number of models fitted to maintain an experiment-wise error rate of α. However, other authors (Fienberg, 1977) prefer to test each model at the nominal α level. Second, as illustrated above, a number of different models can be fit in a given computer run; and multiple computer runs are then used to analyze the data. Most of the log-linear model literature follows this exploratory data analysis approach. Bock (1975), however, recommends a confirmatory approach in which the models of interest are established prior to data collection and then only these models are tested. Unfortunately, this latter approach requires a level of understanding of the factors defining the contingency table and their associations that may not exist prior to data collection. Consequently, the majority of analyses appearing in the literature fall within the exploratory mode.

Example 3. When interest is in a small subset of items rather than all items in a test, the log-linear model approach permits analysis of the test scores at a finer level of detail. Let it be assumed that a population can be divided into two groups, say A and B, on the basis of their instructional programs and that the responses of these examinees to a subset of three items is of interest. For example, it may be suspected that the set of items yields different results as a function of the instructional program used in each group. Since each of the three items has only two possible outcomes, there are 2^3 possible item response patterns, although there are only four possible test scores. Thus, rather than simply comparing total score distributions, the two groups can be compared with respect to the frequency of occurrence of the eight possible item response patterns. A possible contingency table, based upon hypothetical data, is shown in Table 5. Again, the hypothesis of interest is whether the group classification, Factor 1, is independent of the item response patterns, Factor 2. The corresponding model of interest is simply [1], [2] and it yielded a residual $G^2 = 7.16$ with 7 *df* and $p = .413$. Thus, the

Table 5
Frequency of Item Response Patterns for Two Instructional Groups Responding to Three Items

	Pattern								
Group	000	001	010	011	100	101	110	111	Number in Group
A	31	16	16	47	16	47	47	94	312
B	18	9	9	15	9	27	39	55	184

two groups do not differ with respect to the frequency with which the eight item response patterns occurred.

Due to the exponential relationship between the number of items and the number of item response patterns (2^n), this analysis is practical only for a small number of items. Fortunately, this usually holds when there is consideration of a single behavioral objective or a specific instructionally related topic where 3 to 5 items are often used to measure the student's level of performance.

Example 4. An interesting subanalysis of the data in Table 5 can also be performed. From each group, those examinees can be extracted who have a raw score of two which corresponds to the item response patterns, 011, 101, and 110. The question of interest is whether these two subgroups differ with respect to the composition of their common test score. Such an analysis would be of interest to those wishing to examine the possibility of item bias, since if the two groups matched on test score differed in how they obtained that score, item bias might be suspected within the set of items. The appropriate data has been extracted from Table 5 and is presented in Table 6. The log-linear model is simply that for the hypothesized independence of the row [1] and column [2] factors and yielded a $G^2 = 15.26$ with 2 *df*, which was significant at $\alpha = .05$. Thus, the two groups differed with respect to how they obtained a score of two; and for persons having this score, there may be a bias among the three items. An analysis of this type could also be performed for those examinees in the two groups having a score of one.

Table 6
Frequency of Item Response Patterns for Two Groups Matched on the Basis of a Raw Score of Two

	Pattern			
	1	2	3	Number
Group	011	101	110	in Group
A	47	47	47	141
B	15	27	39	97

A number of the usual item bias techniques (see Ironson & Subkoviak, 1979) analyze one item at a time and compare the performance of two or more groups over the entire score scale. The present log-linear procedures can be used to investigate item bias in two ways. First, for sets of items known to be biased, the procedures can be used to determine the score levels at which there is an association between group membership and item response patterns. Second, they can be used to identify sets of items that may be individually unbiased but whose response patterns are associated with group membership only at particular score levels. In both cases the methodology enables the researcher to ascertain the association of group membership with the item response patterns yielded by sets of items rather than only for single items. In addition, it allows this to be done within particular score levels that may be of interest.

An Application of Log-Linear Models to the Results of Criterion-Referenced Tests

Within the context of modern instructional programs, criterion-referenced tests often are used to make a variety of instructionally related decisions. Typically, a criterion-referenced test is used as a unit post-test to determine if a student has mastered the content covered by the unit. If the obtained test score exceeds a criterion level, the student proceeds to the next unit within the curricular plan (mastery). If the obtained test score does not exceed the criterion level, (nonmastery) is indicated, and the student receives remedial procedures and upon completion is administered a post-test again. The relationship of achieving mastery to other variables, both instructional and student related, are of interest to teachers and to curriculum designers. Again, the log-linear model approach provides a very useful vehicle for exploring the interrelationships among such variables.

Example 5. Assume that 37 pupils have been divided into mastery and nonmastery groups on the basis of a criterion-referenced test. The question of interest is whether attaining mastery is associated with two other factors in the instructional setting, namely, the number of curriculum-embedded tests previously failed (0, 1, 2) by the student and the student's reading level (L, M, H). The result would be a 2 × 3 × 3 contingency table where Factor 1 is achievement, Factor 2 is tests failed, and Factor 3 is reading level. A hypothetical set of data for such a situation is shown in Table 7. From a log-linear model point of view, these data have the same structure as that of Example 2.

Table 7
Allocation of Students by Mastery, Prior Attempts and Reading Difficulty

		Reading Level			
Group		L	M	H	
Mastery					
Prior Attempts	0	1	0	6	7
	1	1	2	3	6
	2	3	4	1	8
		5	6	10	21
Non-Mastery					
Prior Attempts	0	3	1	1	5
	1	1	2	1	4
	2	1	1	5	7
		5	4	7	16

The frequency data were analyzed via the 3F program of the BMD/P series and the goodness-of-fit indices are reported in Table 8. Only the model containing all of the first-order associations did not fit the data at a significance level of .05, indicating that there is a second-order association among the three variables. The component χ^2 yielded by the comparison of models [12] [13] [23] and [12] [13], although not significant, suggests that the association of reading level and number of prior tests failed may be of interest. The remaining models all fit the data. Because of this, the suspicion is that other than the second-order association, the data within the contingency table is rather weakly structured.

Table 8
Goodness of Fit Indices for the Data of Table 7

	Residual			Component		
Model	G^2	df	p	G^2	df	p
[123]	0	0	-			
[12][13][23]	9.51	4	.05	9.51	4	.05
[12][13]	12.31	8	.14	2.80	4	.45
[12][23]	9.73	6	.14	.22	2	.88
[13][23]	9.60	6	.14	.09	2	.96
[1][2][3]	12.62	12	.397	3.02	6	.84

Use of Log-Linear Models as an Item Analysis Technique

One of the interesting facets of log-linear models is that they are directly related to the latent trait theory procedures due to Bock (1972) for estimating the difficulty (*b*) and discrimination (*a*) parameters of items that have been dichotomously or nominally scored. These procedures fall within a facet of log-linear models known as logit-linear models. Under this latter approach, the qualitative variables used to define the sampling plan, such as race and school location in Example 2, are called explanatory variables, and the variables used to categorize the examinees' responses, such as test score level in Example 2, are called response variables.

A primary difference between logit-linear and log-linear models is that the latter formulates all the classificatory variables as response variables. Under the logit-linear approach, the marginal totals corresponding to an explanatory variable are considered fixed. In addition, the logit-linear technique uses the multivariate logistic function (see Bock, 1970, 1975) as the vehicle for obtaining the estimated cell frequencies under a given model. When the response variable has only two categories, the multivariate logistic function reduces to the familiar two-parameter logistic function. It is of interest that a number of authors (see Bishop, Fienberg, & Holland, 1975; Everitt, 1977, chap. 5; Fienberg, 1977) introduce logit-linear models via the *r* × 2 contingency table and the two-parameter logistic function. Because the relationship between the algebraic representations of the linear models for a given contingency table under these two approaches are notationally complex, the actual linear models will not be shown in this section. The interested reader should consult Everitt (1977) or Fienberg (1977) for presentation of the mathematics equating the two approaches.

In the present section, two examples will be used to illustrate the relationship between logit-linear models for contingency tables and item parameter estimation under latent trait theory when the examinees' ability scores are known. The first will involve a single item and will be used to demonstrate the basic relationship. The second will show how the parameters of several items can be estimated simultaneously under the logit-linear model approach.

Example 6. Under latent trait theory, the item characteristic curve is the functional relationship between the probability of correct response, p_j, and the ability score θ_j. Using a two-parameter logistic model for the item characteristic curve, then

$$P_j = \frac{1}{1 + e^{-a(\theta_j - b)}} \; . \qquad [11]$$

When the values of θ_j are known, the observed proportion of correct responses to a given item at each ability level can be obtained. The basic task then is to estimate the difficulty and discrimination parameters of the underlying item characteristic curve. In the present example, the four known ability levels will have values of −3, −1, 1, and 3, respectively, and there are n_j = 1,000 examinees at each of the four levels. For didactic purposes, the observed proportions of correct responses will be generated from a two-parameter logistic model (a = 1.0, b = 0.0) for the item characteristic curve. Using these specifications, Equation 11 was evaluated at each ability level and p_j, j = 1, 2, 3, and 4 obtained. Then, the number of correct responses $f_{j1} = p_j n_j$ and the number of incorrect responses $f_{j2} = (1 - p_j)n_j = q_j n_j$ were determined at each ability level. These obtained values can be presented in the form of a two-dimensional array with ability level as rows and item response as columns, as shown in Table 9. If the data were subjected to the usual maximum likelihood estimation procedures for item analysis (Maxwell, 1959), the values of the item parameters *a* and *b* would be recovered.

Under the logit-linear approach, this two-dimensional array is considered to be a 4 × 2 contingency table. The explanatory variable is ability, and the sample design is a stratified sampling plan having four levels. At each of these levels, 1,000 examinees have been sampled. The response design

Table 9
Item Response Data Under a Logistic Model (a = 1.0, b = 0)

Ability Level	Item Response	
	Correct	Incorrect
-3	50	950
-1	269	622
1	731	269
3	950	50

employs a single factor—the item—having two levels, correct and incorrect item response. The logit-linear approach attempts to reproduce the observed cell frequencies by using the two-parameter logistic function as the relationship between the explanatory and response variables.

In the present example, the contingency table data of Table 9 were analyzed via the MULTIQUAL program. The known ability scores were entered via the arbitrary contrast option and the data were the cell frequencies. Since the data were generated from a known item characteristic curve, the parameter estimates obtained, $\hat{a} = .99$, $\hat{b} = 0.0$, were very close to the underlying values. Although the item parameter estimation procedure used in latent trait theory when ability is known and the logit-linear model technique for contingency table analysis approach the data from different frames of reference, they yield equivalent results. The reader is referred to Bock (1975, chap. 8) for the underlying mathematics and a related example.

It is well known (see Maxwell, 1959) that if the observed proportions of correct response are subjected to a logistic transformation, the basic problem becomes one of estimating the parameters of a linear regression line. In this context, ability is the independent variable and the dependent variable is expressed in logits. The contingency data of Table 9 were formulated in terms of linear regression so that the MULTIQUAL program could estimate the parameters. In recent years, the statistical literature has labeled the underlying approach as logistic regression, and it is widely used in fields outside of psychometrics.

Example 7. With the parallelism of item analysis and contingency table analysis in hand, the present example will illustrate how the logit-linear technique can be used to estimate the parameters of several items simultaneously. The explanatory variable will be ability with five known levels. The response design employs three response variables—items—each with two levels, correct and incorrect response. Thus, the response design is a fully crossed 2 × 2 × 2 factorial layout. As was the case in Example 3, the basic data to be collected are the number of examinees possessing each of the eight possible item response patterns. However, in the present example, these patterns are the result of classifying the examinees on the basis of three variables rather than on a single variable.

The data to be analyzed were obtained from the responses of 1,000 examinees to three selected items of the verbal section of the 1975 Scholastic Aptitude Test[1]. Since under the logit-linear approach values of the explanatory variable must be known, it is necessary to obtain ability estimates for each of the examinees. This was done by analyzing the responses of the 1,000 examinees to all 85 items of the test via the LOGIST program (Wood, Wingersky, & Lord, 1976). The obtained ability score estimates were grouped into five intervals having midpoints -2, -1, 0, 1, and 2; and these numbers were used as the known values of the explanatory variable. The frequency of occurrence of correct and incorrect response to the three items was tallied within each of the five ability levels. The observed cell frequencies in the 5 × 2 × 2 × 2 contingency table are reported in Table 10.

[1] These data were provided by the Educational Testing Service with permission of the College Entrance Examinations Board.

Table 10
Observed Frequency of Item Response Patterns
As a Function of Ability

	Item 1	0				1			
	Item 2	0		1		0		1	
	Item 3	0	1	0	1	0	1	0	1
Ability	Pattern	000	001	010	011	100	101	110	111
-2		75	8	39	1	30	1	7	0
-1		60	16	69	11	27	7	26	4
0		21	20	97	43	19	5	52	11
+1		5	8	37	25	8	4	59	29
+2		0	1	14	15	0	6	48	92

Since each item is considered to be a classificatory variable, they can be treated as independent or as associated variables. Assuming the three items are independent of one another, it is possible to simultaneously estimate the discrimination (*a*) and difficulty (*b*) parameters of the three item characteristic curves under a logistic model. The MULTIQUAL program was used to analyze the data of Table 10 and the obtained item parameter estimates, after appropriate rescalings, are given in Table 11. The item parameter estimates yielded by the LOGIST analysis are also shown in Table 11, as they are the values expected if the two approaches agree. With the exception of the estimate of *b* for Item 1, the estimates yielded by MULTIQUAL are reasonably close to the LOGIST results. This general agreement is quite good in light of the rather crude grouping of examinees on the ability scale. Due to the use of item response patterns, the logit-linear approach is not practical when more than a few items are of interest.

In both of the item analysis examples, it was assumed that the ability levels of the examinees were known. Although this is rarely the case, it nonetheless represents the situation within a stage of the joint estimation paradigm (Wood, Wingersky, & Lord, 1976) widely used in latent trait theory. When estimating the item parameters, the current estimates of ability are assumed to be the true values and hence are considered to be known. Thus, a joint estimation paradigm could be established in which the logit-linear model approach was used as the item parameter estimation procedure within each stage. To do so, however, would not be particularly efficient from a computing point of view.

In the present example, ability was considered known and only the item parameters were estimated. The scope of the analysis can be broadened considerably by employing the Rasch model. Under this model the cumbersome patterns of item responses are not needed, as all persons with the same test score receive the same estimated ability. Mellenbergh and Vijn (1981) took advantage of this and formulated the simultaneous estimation of the item difficulty and ability parameters in terms of a log-linear model. They established a three-dimensional contingency table (Items × Test

Table 11
Item Parameter Estimates Yielded By MULTIQUAL
and LOGIST for the Three Items of Table 9

	MULTIQUAL		LOGIST	
Item	a	b	a	b
1	.34	.440	.32	.16
2	.48	-1.066	.54	-1.18
3	.36	1.328	.36	1.30

Score × Item Response) and used a [12] [13] [23] model. The estimated item difficulties and ability estimates yielded by the log-linear procedures were shown to be linear functions of the estimates yielded by the BICAL computer program (Wright & Mead, 1976).

Summary

The data analyses presented above were aimed at illustrating some types of psychometric data analyses that can be handled via log-linear models. Within the context of test norming, the testing of the homogeneity of test score distributions is a useful procedure. Under the log-linear model approach, this can be extended to a finer grain analysis that takes the manner in which the test score was obtained into account as well as incorporating other sampling plan variables. Such procedures have additional potential application in the analysis of criterion-referenced test results, as well as in the area of item and test bias.

A very useful psychometric application of log-linear models is to ascertain the relationships among demographic or instructional and outcome variables. Of immediate interest are analyses where one classificatory variable is mastery of a particular instructional objective, and the remaining variables are characteristics of the instructional setting. Such log-linear analyses provide insight into the relationships among the variables as well as into the composition of these relationships.

Finally, the simultaneous estimation of item parameters for a set of items provides an important theoretical linkage between the procedures of latent trait theory and log-linear models. Overall, the applications presented above show that log-linear models are a versatile data analysis technique and one that has widespread potential applicability in a measurement context.

References

Angoff, W. H. "Scales, norms and equivalent scores." In R. L. Thorndike (Ed.), *Educational measurement.* Washington DC: American Council on Education, 1971.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. *Discrete multivariate analysis: Theory and practice.* Cambridge MA: The MIT Press, 1975.

Bock, R. D. Estimating multinomial response relations. In R. C. Bose (Ed.), *Essays in probability and statistics.* Chapel Hill NC: The University of North Carolina Press, 1970.

Bock, R. D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 1972, *37*, 29-51.

Bock, R. D. *Multivariate methods in behavioral research*. New York: McGraw-Hill, 1975.

Bock, R. D., & Yates, G. *MULTIQUAL*. Chicago IL: National Educational Resources, Inc., 1973.

Dixon, W. J., & Brown, M. B. (Eds.), *BMDP-79: Biomedical Computer Programs P-Series*. Los Angeles: University of California Press, 1979.

Everitt, B. S. *The analysis of contingency tables*. London: Chapman & Hale, 1977.

Fay, R. E., & Goodman, L. A. *ECTA Program: Description for users*. Chicago IL: University of Chicago, Department of Statistics, 1975.

Fienberg, S. E. *The analysis of cross-classified categorical data*. Cambridge MA: The MIT Press, 1977.

Haberman, S. J. *Analysis of qualitative data (Vol. 1)*. New York: Academic Press, 1978.

Haberman, S. J. *Analysis of qualitative data (Vol. 2)*. New York: Academic Press, 1979.

Ironson, G. H., & Subkoviak, M. J. A comparison of several methods of assessing item bias. *Journal of Educational Measurement*, 1979, *16*, 209-226.

Lord, F. M. Test norms and sampling theory. *Journal of Experimental Education*, 1959, *27*, 247-263.

Maxwell, A. E. Maximum likelihood estimates of item parameters using the logistic function. *Psychometrika*, 1959, *24*, 221-227.

Mellenbergh, G. J., & Vijn, P. The Rasch model as a log linear model. *Applied Psychological Measurement*, 1981, *5*, 369-376.

Wood, R. L., Wingersky, M. S., & Lord, F. M. LOGIST: A computer program for estimating ability and item characteristic curve parameters

(Research Memorandum 76-6). Princeton NJ: Educational Testing Service, 1976.

Wright, B., & Mead, R. J. *BICAL: Calibrating items and scales with the Rasch model* (Research Memorandum No. 23). Chicago IL: University of Chicago, Department of Education, Statistical Laboratory, 1977.

Author's Address

Send requests for reprints or further information to Frank B. Baker, Department of Educational Psychology, University of Wisconsin, Madison WI 53706.